6th European Conference on Computational Mechanics (ECCM 6)
7th European Conference on Computational Fluid Dynamics (ECFD 7)
1115 June 2018, Glasgow, UK

SYMBOLIC REGRESSION OF ALGEBRAIC STRESS-STRAIN RELATION FOR RANS TURBULENCE CLOSURE

Martin Schmelzer[1*], Richard P. Dwight[1] and Paola Cinnella[2]

[1] Aerodynamics Group
Faculty of Aerospace Engineering
Delft University of Technology
Kluyverweg 1, 2629 HS Delft, The Netherlands,
m.schmelzer@tudelft.nl and r.p.dwight@tudelft.nl

[2] Laboratoire DynFluid
Arts et Métiers ParisTech
151 Boulevard de l'Hospital, 75013 Paris, France
paola.cinnella@ensam.eu

Key words: Turbulence Modelling, RANS, Explicit Algebraic Reynolds-stress Models, Deterministic Symbolic Regression, Regularised Least-Square Regression, Machine Learning

Abstract. In this work recent advancements are presented in utilising deterministic symbolic regression to infer algebraic models for turbulent stress-strain relation with sparsity-promoting regression techniques. The goal is to build a functional expression from a set of candidate functions in order to represent the target data most accurately. Targets are the coefficients of a polynomial tensor basis, which are identified from high-fidelity data using regularised least-square regression. The method successfully identified a correction term for the benchmark test case of flow over periodic hills in 2D at $Re_h = 10595$.

1 INTRODUCTION

The workhorse in industry to solve the closure problem of the Reynolds-Averaged Navier-Stokes (RANS) equations is still the linear eddy-viscosity (LEV) or Boussinesq hypothesis and corresponding transport models. The lower computational costs compared to high-fidelity approaches, e.g. large-eddy simulation, come at the price of low predictive performance for flows with separation, adverse pressure gradients or high streamline curvature. Explicit Algebraic Reynolds-Stress Models (EARSM) were introduced to lift the predictive fidelity of RANS at similar costs as LEV. Commonly, EARSM are derived by projecting a Reynolds-stress model (RSM) onto a polynomial tensor basis with the intention that the resulting model inherits a part of the predictive fidelity of the underlying RSM.

Recently a new approach based on symbolic regression utilising genetic programming (GP) was introduced to learn the non-linear stress-strain relationship for the anisotropy tensor based on high-fidelity reference data [1, 2]. This data-driven method retains the input quantities used to derive EARSM but replaces the commonly used projection method to find the formal structure of the model by an evolutionary process based on model fitness. In that way it produces models similar to EARSM but with a proven mathematical form to reproduce the data it was trained on. This method has the potential to generate numerically robust models with a high predictive fidelity. Even though the method is non-deterministic it discovers similar expressions for different runs. However, it is not clear if this variability comes from the data or is due to the inherent randomness of GP.

To overcome this characteristic of GP a couple of non-evolutionary methods for symbolic regression have been introduced recently, such as Fast Function Extraction (FFX) [3], Elite Bases Regression (EBR) [4], Sparse identification of nonlinear dynamics (SINDy) [5] or PDE functional identification of nonlinear dynamics (PDE-FIND) [6]. These methods being based on sparsity-promoting linear regression show for a couple of problems similar or better performance and higher convergence rates for high-dimensional problems than GP. Due to their deterministic nature they discover always the same model given input quantities and parameters. By varying the input parameters of the method a hierarchy of models of varying complexity and predictive fidelity can be discovered.

In this work we follow a two-step process. First, we introduce a model-form error term in the constitutive turbulence closure and compute the discrepancy tensor field and regress it onto a polynomial tensor basis, which is used for nonlinear eddy viscosity models. Second, the scalar coefficient fields of the tensor basis are used as targets for the deterministic symbolic regression. Finally, a simulation of the flow over periodic hills in 2D with the identified correction model will be conducted.

2 NON-LINEAR EDDY VISCOSITY MODELS

For RANS-based turbulence modelling commonly the adequate choice of the turbulence model, e.g. $k-\omega$ or $k-\epsilon$, is an essential requirement in order to achieve good predictive performance. However, the constitutive relation between the strain of the velocity field and the Reynolds-stress introduces model-form error which can't be compensated by changing the turbulent transport model leading to error-prone simulations. Commonly used RANS turbulence modelling is based on a linear stress-strain relationship, i.e. Boussinesq approximation, for the nondimensional anisotropic part b_{ij} of the Reynolds-stress τ_{ij}

$$b_{ij} = \frac{\tau_{ij}}{2k} - \frac{1}{3}\delta_{ij} \tag{1}$$

$$b_{ij} = -S_{ij} \tag{2}$$

in which k is the turbulent kinetic energy, τ the turbulent time scale and $S_{ij} = \tau\frac{1}{2}(\partial_j U_i + \partial_i U_j)$ the normalised mean strain rate tensor. The latter represents the symmetric part of the mean velocity gradient tensor $\partial_j U_i$ and the normalised mean rotation tensor $\Omega_{ij} =$

$\tau\frac{1}{2}(\partial_j U_i - \partial_i U_j)$ the corresponding antisymmetric part. In [7] a more general eddy viscosity model was derived based on $b_{ij} = b_{ij}(S_{ij}, \Omega_{ij})$ as a linear combination of ten base tensors

$$b_{ij}(S_{ij}, \Omega_{ij}) = \sum_{n=1}^{10} T_{ij}^{(n)} \alpha_n(I_1, ..., I_5), \tag{3}$$

in which the coefficients α_n are function of five invariants $I_1, ..., I_5$. The first four base tensors T_{ij}^n and two invariants I_m read

$$T_{ij}^1 = S_{ij},\ T_{ij}^2 = S_{ik}\Omega_{kj} = \Omega_{ik}S_{kj},$$
$$T_{ij}^3 = S_{ik}S_{kj} - \frac{1}{3}\delta_{ij}S_{mn}S_{nm},\ T_{ij}^4 = \Omega_{ik}\Omega_{kj} - \frac{1}{3}\delta_{ij}\Omega_{mn}\Omega_{nm}, \tag{4}$$
$$I_1 = S_{mn}S_{nm},\ I_2 = \Omega_{mn}\Omega_{nm}. \tag{5}$$

The finite number of the base tensors can be attributed to the Cayleigh-Hamilton theorem: Any higher order products of the two tensors S_{ij} and Ω_{ij} can be represented by a linear combination of this tensor basis. Given the base tensors the identification of the functional form of the coefficients $\alpha_n(I_1, ..., I_5)$ is the essential step to build a nonlinear eddy-viscosity model. Classical methods to identify the functional forms are based on projecting RSM onto the polynomial basis [7, 8]. In the following we will derive these functions directly from data using deterministic symbolic regression.

3 DEFINITION OF TARGETS FOR SYMBOLIC REGRESSION

Following [2] we introduce an additive term b_{ij}^{Δ} to compensate the model-form error of the Boussinesq approximation

$$\tau_{ij} = -2\nu_t S_{ij} + \frac{2}{3}k\delta_{ij} + 2kb_{ij}^{\Delta}. \tag{6}$$

While the terms τ_{ij}, k and S_{ij} are available from databases of high-fidelity LES or DNS simulations, the eddy viscosity ν_t can be identified by passively solving a turbulence model, e.g. $k-\omega$, for a given velocity field U_i and a modified production term P_k [2]. Thus, for a given turbulence model and test case the model-form error b_{ij}^{Δ} can be identified and serves as our primal target quantity. Following the rationale of nonlinear eddy viscosity models, we use the tensor basis to find functional models representing the model-form error. We first identify corresponding coefficients fields α_n by minimising the $l2$-norm between the model-form error b_{ij}^{Δ} and the tensor basis independently in each mesh point k

$$\alpha_{n,k} = \underset{\hat{\alpha}_{n,k}}{\arg\min} \left(\left\| \sum_{n}^{N} \hat{\alpha}_{n,k} T_{ij,k}^{(n)} - b_{ij,k}^{\Delta} \right\|_2^2 + \lambda_\alpha \left\| \hat{\alpha}_{n,k} \right\|_2^2 \right). \tag{7}$$

The solution of this optimisation problem is [9]

$$\alpha_{n,k} = (M_{nm,k} + \lambda_\alpha \delta_{nm})^{-1} s_{n,k}, \tag{8}$$

with

$$M_{nm,k} = \begin{bmatrix} T^{(1)}_{ij,k}T^{(1)}_{ij,k} & T^{(2)}_{ij,k}T^{(1)}_{ij,k} & \cdots & T^{(N)}_{ij,k}T^{(1)}_{ij,k} \\ & T^{(2)}_{ij,k}T^{(2)}_{ij,k} & \cdots & T^{(N)}_{ij,k}T^{(2)}_{ij,k} \\ & & \ddots & \vdots \\ symm & & & T^{(N)}_{ij,k}T^{(N)}_{ij,k} \end{bmatrix}, \quad s_{n,k} = \begin{bmatrix} b^{\Delta}_{ij,k}T^{(1)}_{ij,k} \\ b^{\Delta}_{ij,k}T^{(2)}_{ij,k} \\ \vdots \\ b^{\Delta}_{ij,k}T^{(N)}_{ij,k} \end{bmatrix}, \quad \alpha_{n,k} = \begin{bmatrix} \alpha_{1,k} \\ \alpha_{2,k} \\ \vdots \\ \alpha_{N,k} \end{bmatrix}. \tag{9}$$

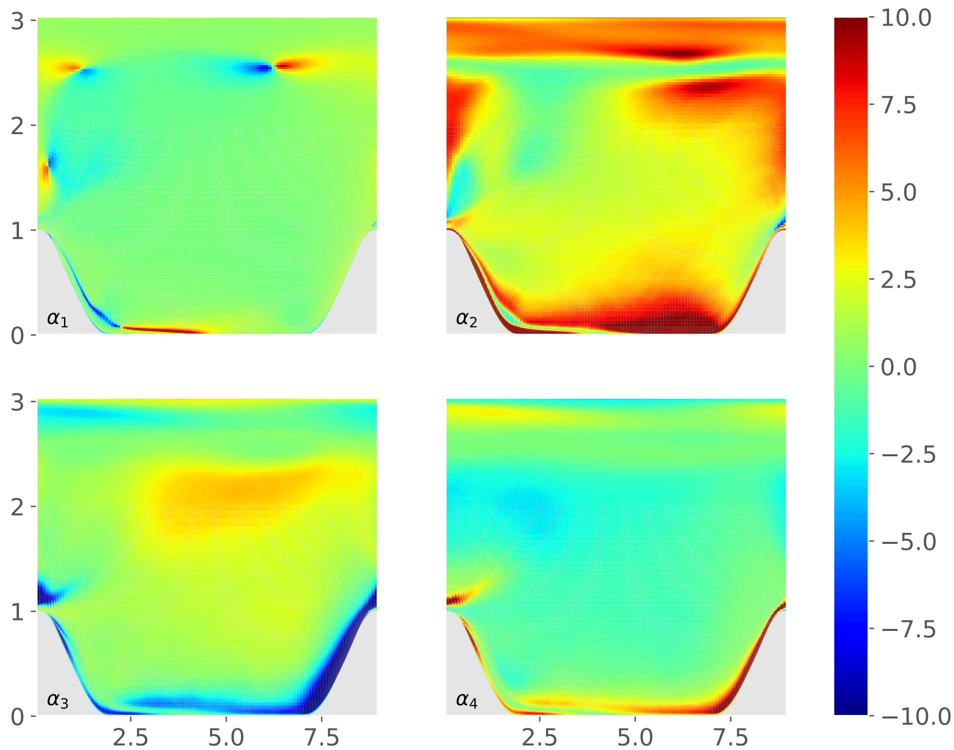

Figure 1: Coefficient fields α_n with $n \in [1, 2, 3, 4]$ for flow over periodic hills in 2D at $Re_h = 10595$ [10].

Ordinary least-square regression, i.e. $\lambda_\alpha = 0$, leads to unphysical behaviour of large differences of the parameter values for neighbouring mesh points and an overall high norm of the scalar coefficient fields. While the identified coefficients lead to perfect reconstruction of the target $b^{\Delta}_{ij,k}$, the coefficients are of low practical value for physical interpretation and symbolic regression. An active regularisation parameter $\lambda_\alpha > 0$ reduces the sensitivity of the inversion to noise in the system and smoothens the resulting scalar coefficient fields $\alpha_{n,k}$ spatially. However, this also introduces bias which increases the reconstruction error. Therefore a parameter study needs to be conducted to identify suitable values for λ_α. For the test case of flow over periodic hills in 2D at $Re_h = 10595$ presented in this work the regularisation parameter was set to $\lambda_\alpha = 0.0001$, which gave an acceptable global mean-squared reconstruction error of $\epsilon_{MSE} = 0.002$. The resulting coefficient fields α_n for four base tensors are presented in Figure 1.

4 DETERMINISTIC SYMBOLIC REGRESSION

The deterministic symbolic regression constructs an over-complete library of possible nonlinear candidate functions and identifies the important ones by adopting a sparsity constrain. Since the general form of our main target depends on the number of base tensors, we will conduct a single symbolic regression for each $\alpha_{n,k}$ with $n \in [1, ..., 4]$. Given a set of simple input features, i.e. the invariants I_1 and I_2, we build a library matrix B_{mk} of nonlinear combinations of these simple input features. As a starting point we build the candidates by taking the simple input features to a certain power, I_m^p with $p \in [0.5, 1, 2]$, and by taking the product between each of the resulting functions, which leads to functions of the form e.g. $I^{0.5}$ or $I_1 \cdot I_2^2$. Also more complex operations can be used leading to a larger library. Given the library of candidates, the optimisation problem of the symbolic regression can be stated as

$$\Theta_m^{(n)} = \arg\min_{\hat{\Theta}_m} \left\| B_{mk}\hat{\Theta}_m - \alpha_{n,k} \right\|_2^2 + \lambda \left\| \hat{\Theta}_m \right\|_q , \tag{10}$$

in which the vector $\Theta_m^{(n)}$ needs to be identified. The target is a specific $\alpha_{n,k}$ represented by a column vector of size k containing all values of $\alpha_{n,k}$ from each mesh point. The regularisation term using norm $q = 1$ (LASSO) or $q = 2$ (RIDGE) acts to increase the sparsity of $\Theta_m^{(n)}$, i.e. increasing the number of zeros in order to turn off the corresponding base functions [5, 6]. The result is the vector $\tilde{\alpha}_{n,k} = B_{mk}\Theta_m^{(n)}$, i.e. the discovered model, in which $\Theta_m^{(n)}$ indicates which base functions are retained by assigning a non-zero value to them. Based on mean LES data for the flow over periodic hills in 2D at $Re = 10595$ [10], using four base tensors $N = 4$ and two simple input features I_1 to I_2, from which the library matrix is build, an identified model reads

$$\begin{aligned} M_{PH} := {} & (5.02 \cdot I_2 + 2.83 \cdot I_1^{0.5})\, T_{ij}^{(1)} \\ & + (57.38 \cdot I_1^{0.5} - 152.98 \cdot I_1)\, T_{ij}^{(2)} \\ & - (43.66 \cdot I_2 + 42.16 \cdot I_1^{0.5} \cdot I_2 + 9.48 \cdot I_1^{0.5})\, T_{ij}^{(3)} \\ & + (7.20 \cdot I_1^{0.5} - 33.52 \cdot I_1)\, T_{ij}^{(4)}. \end{aligned} \tag{11}$$

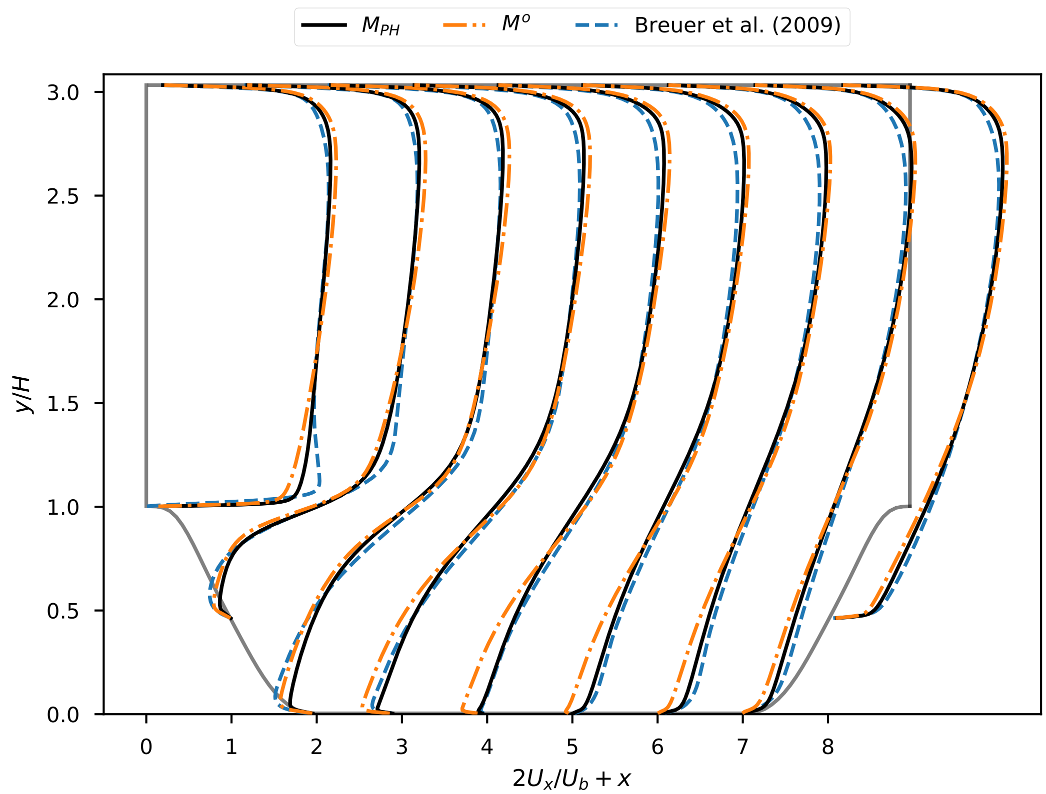

Figure 2: Stream-wise velocity profiles at several downstream locations for the flow over periodic hills at Re=10595.

The model uses per base tensor two to three candidate functions and is therefore of similar complexity as models derived with GP in [11, 12]. In Figure 2 the stream-wise velocity at several cross-sections is displayed for M_{PH} and for a linear baseline simulation M^o using the linear $k - \omega$ model in comparison to the reference data [10]. Overall the velocity prediction of M_{PH} is closer to the reference data and especially the circulation zone including the reattachment of the flow behind the hill is predicted more accurately.

5 CONCLUSION

In this paper it was shown that deterministic symbolic regression based on sparsity-promoting regularisation can be used to identify nonlinear correction terms for the turbulent stress-strain relation. The method was applied to the flow over periodic hills in 2D, a challenging benchmark test case for RANS, and successfully identified a correction term which led to a better velocity prediction. Also the mathematical complexity is similar to models present in literature derived or identified with other means. This promising first result is limited by the fact that both the identification of the model as well as the prediction is done on the same test case. Further research will both focus on identifying models for different test cases and Reynolds-numbers as well as validation of the predictive performance of the models when applied to other test cases.

REFERENCES

[1] Weatheritt, J. and Sandberg, R., "A novel evolutionary algorithm applied to algebraic modifications of the RANS stressstrain relationship," *Journal of Computational Physics*, Vol. 325, 2016, pp. 22–37.

[2] Weatheritt, J. and Sandberg, R. D., "The development of algebraic stress models using a novel evolutionary algorithm," *Flow, Turbulence and Combustion*, 2017.

[3] Mcconaghy, T., "FFX: Fast, Scalable, Deterministic Symbolic Regression Technology," *Genetic Programming Theory and Practice IX*, 2011, pp. 235–260.

[4] Chen, C., Luo, C., and Jiang, Z., "Elite Bases Regression: A Real-time Algorithm for Symbolic Regression," Vol. 1, No. 2, 2017.

[5] Brunton, S. L., Proctor, J. L., and Kutz, J. N., "Discovering governing equations from data: Sparse identification of nonlinear dynamical systems," Vol. 113, No. 15, 2015, pp. 3932–3937.

[6] Rudy, S. H., Brunton, S. L., Proctor, J. L., and Kutz, J. N., "Data-driven discovery of partial differential equations," *Science Advances*, 2017.

[7] Pope, S. B., "A more general effective-viscosity hypothesis," *Journal of Fluid Mechanics*, Vol. 72, No. 2, 1975, pp. 331–340.

[8] Apsley, A. D. and Leschziner, M. A., "A New Low-Re Non-Linear Two-Equation Turbulence Model for Complex Flows," *International Journal of Heat and Fluid Flow*, Vol. 19, No. 3, 1998, pp. 209–222.

[9] Aster, R., Borchers, B., and Thurber, C. H., *Parameter Estimation and Inverse Problems*, Elsevier Inc., 2nd ed., 2009.

[10] Breuer, M., Peller, N., Rapp, C., and Manhart, M., "Flow over periodic hills - Numerical and experimental study in a wide range of Reynolds numbers," *Computers and Fluids*, Vol. 38, No. 2, 2009, pp. 433–457.

[11] Weatheritt, J., Sandberg, R., and Lozano-Durán, A., "Reynolds Stress Structures in the Hybrid RANS/LES of a Planar Channel," *Journal of Physics: Conference Series*, Vol. 708, No. May, 2016, pp. 012008.

[12] Weatheritt, J., Sandberg, R. D., Laskowski, G., and Michelassi, V., "Machine Learning For Turbulence Model Development Using A High-Fidelity HPT Cascase Simulation," *Turbomachinery Technical Conference and Exposition GT2017*, 2017, pp. 1–12.