EPJ Web of Conferences **140**, 12006 (2017) DOI: 10.1051/epjconf/201714012006
Powders & Grains 2017

Testing Occam's razor to characterize high-order connectivity in pore networks of granular media: Feature selection in machine learning

Joost van der Linden[1], *Antoinette* Tordesillas[2,⋆], and *Guillermo* Narsilio[1]

[1] *Department of Infrastructure Engineering, The University of Melbourne, Australia*
[2] *School of Mathematics and Statistics, School of Earth Sciences, The University of Melbourne, Australia*

Abstract. A perennial challenge for the characterization and modelling of phenomena involving granular media is that the internal connectivity of, and interactions between, the pores and the particles exhibit hallmarks of complexity: multi-scale and nonlinear interactions that lead to a plethora of patterns at the mesoscale, including fluid flow patterns that ultimately render a permeability of the granular media at the macroscale. A multitude of physical parameters exist to characterize geometry and structure, including pore/particle shape, volume and surface area, while a rich class of complex network parameters quantifies internal connectivity of the pore and particles in the material. A large collection of such variables is likely to exhibit a high degree of redundancy. Here we demonstrate how to use feature selection in machine learning theory to identify the most informative and non-redundant, yet parsimonious set of features that optimally *characterizes* the interstitial flow properties of porous, granular media, e.g., permeability, from high resolution data.

1 Introduction

Porous, granular materials are complex systems that embody rich patterns and dynamics. Applications involving physical flow, such as hydrocarbon recovery and geothermal energy, rely on estimations of the (coupled) hydraulic, thermal and mechanical material properties. For instance, significant experimental evidence has shown that transport properties of materials, such as permeability, are strongly influenced by the presence of concentrated zones of inelastic deformation such as shear bands, compaction bands, fractures and joints (e.g., [1]). Localized deformation can strongly influence flow pathways, potentially becoming either barriers or conduits for flow depending on the attendant evolution of local pores (e.g., see [3] and references therein). Many of these patterns occur naturally in frictional soft materials, especially in rocks and soil, due to material discontinuities and heterogeneity [2]. Microstructural grain rearrangements can alter permeability not only through changes in the geometry and size distribution of individual pores but also their connectivity [3].

Almost invariably, these complex and iterrelated processes can only be captured in a high-dimensional multivariate parameter space. Such a dataset is generated in this work to characterize permeability using the data-driven framework introduced in [4]. The framework fuses proven finite-element and discrete-element methodology with modern advances in statistics (machine learning) and complex systems (complex networks), towards a data-driven 3D analysis of multiscale and nonlinear phenomena in granular, porous media. Furthermore, the framework can be used to address the coupled evolution of the solid grain and interstitial pore phases through a study of two classes of interdependent networks in a single platform, *viz.* one that represents the grain contacts while the other represents the pores, thus advancing the approach in [3] for planar systems.

As shown in [4], the high-dimensional parameter space of variables related to permeability is likely to exhibit a high degree of redundancy. Occam's razor dictates all but the *most relevant* and *least redundant* of these variables should be retained to explain a given phenomenon of interest. To this end, a ranking of the variables in order of relevance and redundancy is crucial for predictive modelling and ultimately control. We apply two such algorithms in this paper, quantify the redundancy, and highlight several highly relevant parameters for the permeability.

2 Methods

The methods and dataset used here are based on the methodology outlined in [4]. In this paper, we summarize the corresponding framework. For a comprehensive discussion of all model equations, parameters and assumptions involved, we refer to the aformentioned study. The employed data generation process is shown in Figure 1.

2.1 Discrete element modeling

Our discrete-element model is a simple proxy for sands and sedimentary rocks (i.e., Ottawa sand and sandstone) [2, 5, 6]. Batches of 400 soft spheres are dropped in a rectangular container with periodic boundary conditions for

⋆e-mail: atordesi@unimelb.edu.au

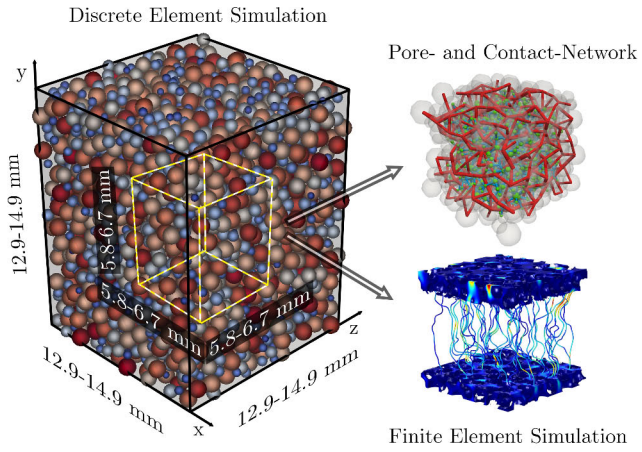


Figure 1. A discrete element simulation is performed first, to generate a compressed packing of spheres. From this packing, a subsample is selected to generate a weighted pore network and weighted contact network. Fluid flow is simulated in this subsample in a finite element simulation, to compute the permeability.

the vertical walls. Each batch is released at equal height from the top of the existing packing, and allowed to settle under gravity, simulating the air pluviation method of sample preparation. After dropping 4000 particles, the resulting packing is subjected to an isotropic confining pressure, leading to changes in porosity. To generate a large (synthetic) dataset, this process of gravity deposition followed by triaxial compression is repeated 1536 times, each time randomly selecting the particle radius distribution ($U(0.5-\alpha, 0.5+\alpha)$ mm, $\alpha \in U(0.0, 0.3)$), friction angle ($\theta \in U(5.7°, 31°)$) and confining pressure (10^n Pa, $n \in U(5,7)$). The Young's modulus is set at the relatively low value of 10^8 Pa. Although this choice is lower than the usual quartz value, it does allow for a simple approximation of a wide range of porosities. At high compression rates, the particle overlap models sintered materials such as sandstones [2].

2.2 Finite element modeling

A representative element volume (REV) is extracted from the center of each packing. Next, we run a finite element simulation for each REV, imposing a pressure difference of 1 Pa over the top and bottom boundaries and solving the governing Navier-Stokes equations to simulate incompressible fluid flow. The permeability of each assembly is computed by averaging the vertical velocity component over the top and bottom boundaries, and subsequently averaging as in [7].

2.3 Complex networks

In van der Linden et al. [4] we constructed both a contact network and a pore network. For details of the pore network construction approach, we refer to the aforementioned paper and our upcoming work [8]. In this study, for simplicity, we omit the contact network results. Every edge in the pore network is weighted with the throat conductance, derived from the Hagen-Poiseuille equation in a simple tube model of the corresponding (adjacent) pores (see [4], Appendix 1).

We compute a range of complex network parameters for the pore network. These parameters, in various ways, capture the connectivity and topology in the network. For a detailed overview, we refer to the seminal work by Newman [9], and for a brief overview of the equations employed in this work, see [4]. For each node, we compute the *degree* as the number of adjacent edges (also known as the coordination number), and the *weighted degree* as the sum of the weights of adjacent edges. The *network density* is the fraction of the actual number of edges over the number of edges if the network were to be fully connected. Furthermore, define $d(i, j)$ as the length of the (*shortest*) path from node i to node j that minimizes the summed weights of the traversed edges. The *network diameter* is the length of the 'longest' shortest path between all pairs of nodes. We also compute the *node betweenness centrality* for a particular node as the fraction of all shortest paths in the network passing through this node. Similarly, we compute *edge betweenness centrality* as the fraction of all shortest paths passing through a particular edge. As such, betweenness centrality captures the 'importance' of a node/edge for the flow through the network. Lastly, we compute the *closeness centrality*, which can be understood as the 'centrality' of a particular node, meaning that the node is well-connected to the rest of the network. Averaged over all nodes, high closeness centrality indicates a well-connected network (i.e. large pores/throats) for which we hypothesize a higher value of the permeability.

2.4 Features

Having generated the required dataset, the final step in our framework is to collate all physical (1-13) and network (14-20) parameters in a 'feature set', as shown in Table 1. Table 1 can be interpreted as our 'initial guess' of 20 features that characterize the permeability to some degree. Connectivity is highly relevant for the permeability [10], yet none of the traditionally used physical features (1-13) capture this aspect. This is why we include the network features (14-20). For any of the distributions (bracket notation in Table 1), we compute the average (denoted with μ) to obtain a single parameter. In the analysis, we consider the natural logarithm of $p, [T]_K, [B]_{A_c}, \mu[G^p]_{k_w}$ and $\mu[G^p]_{C^c}$, because these features span multiple orders of magnitude.

2.5 Feature selection

For each packing, we obtain the values of the parameters listed in Table 1 (the features) and the value of the permeability obtained from the finite element simulations. Given such a high-dimensional dataset of features and a 'target variable' (permeability), we can apply machine learning techniques and, in particular, feature selection, to (1) uncover the most 'important' features for the permeability,

Table 1. Feature set. The notation $[X]_a$ is used to denote a distribution of parameter a on entity X (T for throat, P for pore, B for particle, G for graph).

#	Notation	Attribute	Units
1	e	global void ratio	
2	p	confining pressure	[Pa]
3	ssa	specific surface area	[m^{-1}]
4	$[T]_K$	throat conductance	[$m^3Pa^{-1}s^{-1}$]
5	$[T]_e$	throat void ratio	
6	$[P]_e$	pore void ratio	
7	$[B]_{A_c}$	particle contact area	[m^2]
8	$[T]_V/[P]_V$	throat/pore volume ratio	
9	$[P]_{A_s}$	pore surface area	[m^2]
10	α	grain size range	[m]
11	c_u	coefficient of uniformity	
12	c_c	coefficient of curvature	
13	θ	friction angle	[deg]
14	G^p_ρ	network density	
15	G^p_D	network diameter	[m^{-3}Pa s]
16	$[G^p]_\kappa$	degree	
17	$[G^p]_{\kappa_w}$	weighted degree	[m^{-3}Pa s]
18	$[G^p]_{C^B_{edge}}$	edge betwnness centrality	
19	$[G^p]_{C^B_{node}}$	node betwnness centrality	
20	$[G^p]_{CC}$	closeness centrality	[$m^3Pa^{-1}s^{-1}$]

(2) eliminate redundant features, and (3) predict the target variable for new, previously unseen packings. In this paper, we focus on the first and second task. For a discussion of the third task, refer to [4] for preliminary results.

Two feature selection algorithms are considered in this paper. The *minimum redundancy, maximum relevance* (mRMR) [11] feature selection algorithm minimizes redundancy in the set of selected features, while simultaneously maximizing dependency between selected features and the target variable. Relevance and redundancy is calculated in terms of mutual information, a measure of mutual dependence between two variables. The mRMR algorithm uses an incremental search procedure, in which the most relevant feature is selected first. In subsequent iterations, the most relevant feature that also minimizes redundancy with the previously selected feature(s) is added to the feature set. The mRMR method is myopic, meaning that conditional dependencies between features are ignored.

In contrast, the *RReliefF* [12] feature selection method estimates a feature's importance while taking conditional dependencies into account (i.e. non-myopic). RReliefF does so by rewarding features that separate dissimilar values of the target variable, and penalizing features that separate similar values of the target variable. A score is assigned based on the degree to which a feature uniquely explains the target variable. Consequently, RReliefF implicitely penalizes redundancy by dividing up the score awarded for explaining a corresponding portion of the target variable among the redundant features.

3 Results

Before considering the correlations between features and the permeability, it is instructive to analyze the degree of redundancy *within* the feature set. Figure 2 shows the linear correlation matrix, providing an overview of the inter-feature correlations. Broadly, three groups can be dis-

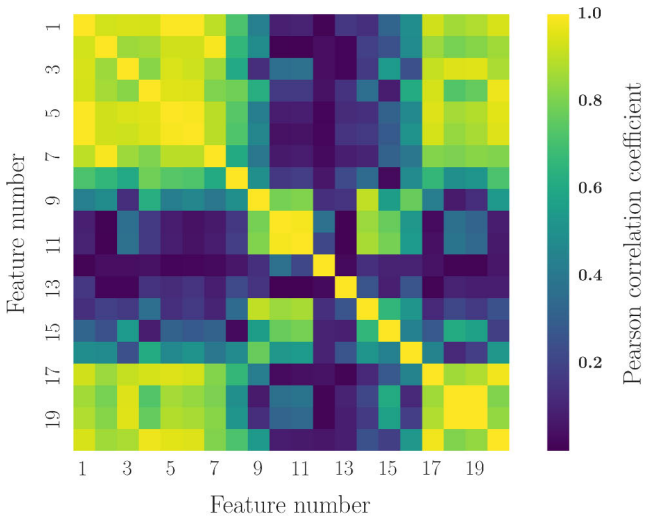
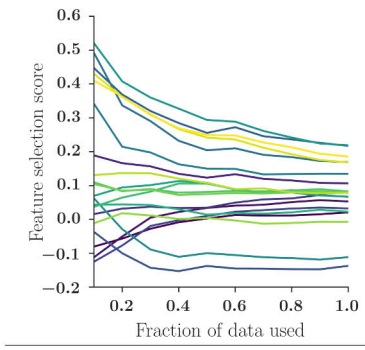


Figure 2. Inter-feature correlations.

tinguished. The first group of features (1-9) are strongly inter-correlated, with an average inter-correlation coefficient of 0.77. These features all relate to the degree of compression and the void ratio. Higher compression results in lower global and local void ratio, as well as higher average inter-particle contact area. Features 1-9 are weakly correlated with the second group: features 10-15. The latter group of features is less correlated with the sample compression and the resulting effect on the void ratio. The third group of features, 16-20, is both inter-correlated (0.66 on average) and correlated with the compression-related features in the first group (1-9). This demonstrates that the connectivity in the samples is broadly interlinked with the degree of compression.

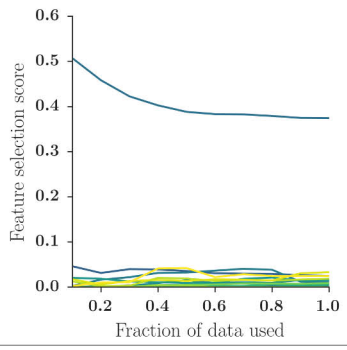
A limitation of using the Pearson correlation coefficient to measure inter-feature correlations is that it does not pick up on non-linear relationships. The coefficients of curvature and uniformity, for example, have a clearly defined non-linear (parabolic) relationship, yet the Pearson correlation coefficient is only 0.20. We do find noteworthy linear correlations, though, between the average pore surface area and the network density (0.90), and between the coefficient of uniformity and the network diameter (0.79). These correlations warrant further investigations.

Having observed clear redundancy in the feature set, we apply the mRMR and RReliefF feature selection methods to apply 'Occam's razor' and analyze the full feature set from Table 1 to identify the most relevant (for the permeability) and non-redundant feature set. For ReliefF, the scores converge as the fraction of data approaches 1.0. Both the local (pore and throat) void ratio and global void ratio are represented in the top 5, consistent with well-known importance of the void ratio for the permeability. A high RReliefF score indicates high relevance of the feature for the permeability and/or strong conditional dependence on the permeability, given other features. The confining



RReliefF

Final ranking				Score
1	log	μ	$[T]_K$	.2187
2	log	μ	$[P]_e$	.2172
3			e	.1856
4	log	μ	$[G^p]_{CC}$	.1691
5		μ	$[T]_e$	.1681
6	log	μ	$[G^p]_{k_w}$	.1347
7			ssa	.1064
8		μ	$[G^p]_{C^B_{edge}}$	.0826
9			G^p_ρ	.0790
10		μ	$[T]_V/[P]_V$	.0788
11		μ	$[G^p]_{C^B_{node}}$	.0780
12		μ	$[P]_{A_r}$	.0671
13			θ	.0665
14			c_u	.0531
15			α	.0323
16		μ	$[G^p]_k$	.0226
17			c_c	.0205
18			G^p_D	−.0073
19	log		p	−.1116
20	log	μ	$[B]_{A_c}$	−.1373



mRMR

Final ranking				Score
1	log	μ	$[G^p]_{CC}$	.3744
2		μ	$[T]_e$	.0329
3	log	μ	$[P]_e$	.0248
4			e	.0247
5		μ	$[T]_V/[P]_V$	.0185
6	log		p	.0172
7	log	μ	$[T]_K$	.0154
8	log	μ	$[G^p]_{k_w}$	.0108
9		μ	$[G^p]_{C^B_{node}}$	.0088
10		μ	$[G^p]_k$	.0066
11		μ	$[P]_{A_r}$	.0048
12	log	μ	$[B]_{A_c}$	.0040
13			θ	.0025
14			ssa	.0019
15			G^p_ρ	.0000
16			α	.0000
17		μ	$[G^p]_{C^B_{edge}}$	.0000
18			G^p_D	.0000
19			c_u	.0000
20			c_c	.0000

Figure 3. (Figures) Convergence of the selection scores, for RReliefF (left) and mRMR (right). (Tables) Final ranking and scores when 100% of the data is used, for RReliefF (left) and mRMR (right).

pressure is strongly correlated with the various void ratios (see Fig. 2), yet p appears at the bottom of the RReliefF ranking. We conclude that, according to RReliefF, p is neither able to explain variance in the permeability by itself, nor does it explain the permeability well when conditioned on other features. A similar conclusion applies to the coefficients of uniformity c_u and curvature c_c, friction angle θ and grain size range α, although we do show in [4] that manually picked combinations of certain features (such as ssa and c_u) can explain part of the variance in permeability.

Recall that the mRMR algorithm follows an incremental search procedure. The convergence of the feature selection scores in Figure 3 for mRMR shows that after picking the first feature ($[G^p]_{C^C}$) subsequent choices are highly penalized for redundancy. In other words, the pore network closeness centrality is highly relevant (in terms of mutual information) for the permeability, and none of the other features are able to explain the permeability much further.

4 Conclusion

The data-driven framework is applied to generate and analyze a set of physical and connectivity descriptors of the permeability. Using feature selection algorightms, we (objectively) show that the initially proposed feature set is highly redundant and can be reduced to a few void ratio descriptors and the pore network closeness centrality. The latter, in particular, captures both the connectivity and (through the conductance weighting) the pore- and throat-geometry of the packings. Upcoming research focuses on characterizing the thermal conductivity and pore space evolution in shear zones [13] using this general framework.

Acknowledgements

The authors acknowledge the support of the Australian Research Council (FT140100227, DP120104759), the U.S. Air Force (AFOSR 15IOA059), and the U.S. Army Research Office (W911NF-11-1-0175, W911NF-15-1-0527).

References

[1] M.A. Etheridge, V.J. Wall, S.F. Cox, R.H. Vernon, Journal of Geophysical Research: Solid Earth **89**, 4344 (1984)

[2] S. Torquato, *Random Heterogeneous Materials: Microstructure and Macroscopic Properties*, Vol. 16 (Springer-Verlag, New York, 2002)

[3] S. Russell, D. Walker, A. Tordesillas, Journal of the Mechanics and Physics of Solids **88**, 227 (2015)

[4] J.H. van der Linden, G.A. Narsilio, A. Tordesillas, Physical Review E **94**, 022904 (2016)

[5] H. Zhu, Z. Zhou, R. Yang, A. Yu, Chemical Engineering Science **63**, 5728 (2008)

[6] G. Narsilio, J. Kress, T. Yun, Computers and Geotechnics **37**, 828 (2010)

[7] G. Narsilio, O. Buzzi, S. Fityus, T. Yun, D. Smith, Computers and Geotechnics **36**, 1200 (2009)

[8] J. van der Linden, A. Sufian, G. Narsilio, A. Russell, A. Tordesillas (2016), unpublished manuscript

[9] M.E.J. Newman, *Networks: an introduction* (Oxford University Press, New York, NY, 2010)

[10] J.T. Fredrich, W.B. Lindquist, International Journal of Rock Mechanics and Mining Sciences **34**, 368 (1997)

[11] H.C. Peng, F.H. Long, C. Ding, IEEE Transactions on Pattern Analysis and Machine Intelligence **27**, 1226 (2005)

[12] M. Robnik-Šikonja, I. Kononenko, Machine Learning **53**, 23 (2003)

[13] A. Tordesillas, J. van der Linden, G. Narsilio (2016), unpublished manuscript