CaTchDes: MATLAB codes for Caratheodory-Tchakaloff Near-Optimal Regression Designs

Len Bos

Department of Computer Science, University of Verona (Italy)

Marco Vianello

Department of Mathematics, University of Padova (Italy)

Abstract

We provide a MATLAB package for the computation of near-optimal sampling sets and weights (designs) for n-th degree polynomial regression on discretizations of planar, surface and solid domains. This topic has strong connections with computational statistics and approximation theory. Optimality has two aspects that are here treated together: the cardinality of the sampling set, and the quality of the regressor (its prediction variance in statistical terms, its uniform operator norm in approximation theoretic terms). The regressor quality is measured by a threshold (design G-optimality) and reached by a standard multiplicative algorithm. Low sampling cardinality is then obtained via Caratheodory-Tchakaloff discrete measure concentration. All the steps are carried out using native MATLAB functions, such as the `qr` factorization and the `lsqnonneg` quadratic minimizer.

Keywords: Near-Optimal Regression Designs, Tchakaloff theorem, Caratheodory-Tchakaloff measure concentration

Email addresses: `leonardpeter.bos@univr.it` (Len Bos), `marcov@math.unipd.it` (Marco Vianello)

Required Metadata

Current code version

Nr.	Code metadata description	Please fill in this column
C1	Current code version	v1.0
C2	Permanent link to code/repository used for this code version	*https://github.com/marcovianello /CaTchDes*
C3	Code Ocean compute capsule	
C4	Legal Software License	GNU/General Public License
C5	Code versioning system used	none
C6	Software code languages, tools, and services used	MATLAB
C7	Compilation requirements, operating environments & dependencies	
C8	If available Link to developer documentation/manual	
C9	Support email for questions	*marcov@math.unipd.it*

Table 1: Code metadata (mandatory)

1. Motivation and significance

The software package `CaTchDes` contains two main MATLAB functions for the computation of *near-optimal sampling sets and weights (designs) for polynomial regression* on discrete design spaces (for example grid discretizations of planar, surface and solid domains). This topic has strong connections with computational statistics and approximation theory: cf., e.g., [1, 2, 3] and the references therein. As a relevant application we may quote for example geo-spatial analysis, where one is interested in reconstructing/modelling a scalar or vector field (such as the geo-magnetic field) on a region with a possibly complex shape, by placing a relatively small sensor network.

In the regression context, optimality has two aspects that are here treated together: the cardinality of the sampling set, and the quality of the regressor (its prediction variance in statistical terms, its uniform operator norm in approximation theoretic terms). Concerning cardinality, a key theoretical tool is the Tchakaloff theorem [4], which in its general version essentially says that for any finite measure there exists a discrete measure that has the same moments up to a given polynomial degree, with cardinality not greater than the dimension of the corresponding polynomial space; cf., e.g., [5].

We briefly recall the statistical notion of optimal design. A *design* is in general a probability measure μ supported on a continuous or discrete compact set X (the design space). In this paper we deal essentially with finite discrete design spaces. Below, we shall denote by $\mathbb{P}_n^d(X)$ the space of d-variate polynomials of total degree not exceeding n and by N_n its dimension.

There are several notions of design optimality. In this work we are mainly interested in G-optimality, that is when the Christoffel polynomial (i.e., the diagonal of the reproducing kernel) has the smallest possible max-norm on X among all designs:

$$\max_{x \in X} K_n^{\mu^*}(x,x) = N_n = \min_{\mu} \max_{x \in X} K_n^{\mu}(x,x) \, , \tag{1}$$

where $K_n^\mu(x,x) = \sum_{j=1}^{N_n} \phi_j^2(x) \in \mathbb{P}_{2n}^d(X)$ and $\{\phi_j\}_{j=1}^{N_n}$ is any μ-orthonormal polynomial basis for degree n. Observe that $\max_{x\in X} K_n^\mu(x,x) \geq N_n$ for any design, since $\int_X K_n^\mu(x,x)\, d\mu = N_n$. This essentially means that a G-optimal design μ^* minimizes both, the maximum prediction variance by n-th degree regression (statistical interpretation), and the uniform norm of the corresponding weighted least-squares operator which has the minimal bound $\sqrt{N_n}$ (approximation theoretic interpretation). In approximation theory, this is also called an optimal measure [6, 2].

The above min-max problem is hard to solve, but by the celebrated Kiefer-Wolfowitz equivalence theorem [7] the notion of G-optimality is equivalent to D-optimality, that is the determinant of the Gram matrix in a fixed polynomial basis is maximal among all designs. This implies that an optimal measure exists, since the set of Gram matrices of probability measures is compact and convex; see, e.g., [6, 8] for a general proof of these facts. By the Tchakaloff theorem, it is then easily seen that an *optimal discrete* measure exists, with $N_n \leq card(supp(\mu^*)) \leq N_{2n}$.

The computational literature on D-optimal designs is quite vast, with a long history and new active research directions, see e.g. [3, 9] and the references therein. A typical approach in the continuous case consists in the discretization of the compact set and then iterative D-optimization over the discrete set. We stress that, due to the convexity of the scalar matrix function $-log(det(\cdot))$, D-optimization in the discrete case is ultimately a convex programming problem, being equivalent to minimizing $-\log(det(V^t D(\mathbf{w})V))$ with the constraints $\mathbf{w} \geq \mathbf{0}$, $\|\mathbf{w}\|_1 = 1$ (where $V = (p_j(x_i)) \in \mathbb{R}^{M \times N_n}$ is the Vandermonde (evaluation) matrix at $X = \{x_i\}$, $1 \leq i \leq M := card(X)$, in a fixed polynomial basis $\{p_j\}$, $1 \leq j \leq N_n$, and $D(\mathbf{w})$ is the diagonal probability weights matrix). We remark that the matrix $V^t D(\mathbf{w}) V$ is equal to the Gram matrix of the polynomial basis $\{p_j\}$, with respect to the discrete measure supported on X with weights $\mathbf{w}$.

2. Software description

Being interested in G-optimality, a relevant indicator is the so-called G-efficiency, namely

$$\theta = N_n / max_{x \in X} K_n^{\mu}(x, x) \tag{2}$$

(the percentage of G-optimality reached). We have pursued the following approach, recently proposed in [10]:

- apply a standard iterative algorithm like Titterington's multiplicative algorithm [11, 12], to get a design $\tilde{\mu}$ with weights $\tilde{\mathbf{w}}$ (i.e., $\tilde{\mu}$ is a discrete measure supported on X with weights $\tilde{w}_i \geq 0$, $1 \leq i \leq M$) possessing a good G-efficiency (say e.g. 95% to fix ideas) in a few iterations;

- compute the Caratheodory-Tchakaloff concentration of the design $\tilde{\mu}$ at degree $2n$, keeping the same orthogonal polynomials and thus the same G-efficiency, with a much smaller support.

We recall that Titterington's multiplicative iteration is simply

$$w_i(k+1) = K_n^{\mu(\mathbf{w}(k))}(x_i, x_i)\, w_i(k)\ ,\ \ 1 \leq i \leq M = card(X)\ ,\ \ k \geq 0\ , \tag{3}$$

starting for example from $\mathbf{w}(0) = (1/M, \ldots, 1/M)$, and is known to converge sublinearly (producing an increasing sequence of Gram determinants) to an optimal design on X; cf., e.g., [12]. Since a huge number of iterations would be needed to concentrate the measure on the optimal support, our approach gives a reasonably efficient hybrid method to nearly minimize both the regression operator norm and the regression sampling cardinality.

Indeed, in the discrete case the Tchakaloff theorem can be stated in terms of the existence of a sparse nonnegative solution to the underdetermined linear system $V^t\mathbf{u} = V^t\tilde{\mathbf{w}}$. Such a solution exists by the celebrated Caratheodory theorem on finite-dimensional conic combinations [13], applied to the columns of V^t. Moreover, it can be conveniently implemented by solving the NNLS (NonNegative Least Squares) problem

$$\min\{\|V^t\mathbf{u} - V^t\tilde{\mathbf{w}}\|_2^2\, ,\, \mathbf{u} \geq \mathbf{0}\} \tag{4}$$

via the Lawson-Hanson active-set iterative method [14], that seeks a sparse solution and is implemented by the basic MATLAB function `lsqnonneg`. It results that the nonzero components of $\mathbf{u}$ determine the Caratheodory-Tchakaloff concentrated support. Let us denote by $\mathbf{u}^*$ the resulting compressed vector of non-zero weights.

This kind of approach to discrete (probability) measure concentration, that can be obtained also via Linear Programming, emerged only recently;

cf., e.g., [15, 16, 17, 18]. We notice that sparsity cannot here be recovered by standard Compressive Sensing algorithms (ℓ^1 minimization or penalization, cf. [19]), since we deal with probability measures and thus the 1-norm of the weights is constrained to be equal to 1.

In the software package `CaTchDes` the near-optimization algorithm above is implemented by the MATLAB function `NORD` (Near-Optimal Regression Design computation), which in turn calls the function `CTDC` (Caratheodory-Tchakaloff Design Concentration). The Vandermonde-like matrix V is constructed using the Chebyshev product basis of the minimal box containing the discrete set X. Both routines automatically adapt to the actual polynomial space dimension, by QR with column pivoting and numerical rank determination for V (this rank gives the numerical dimension of the polynomial space on X). In such a way we can treat cases where X is not determining for the full polynomial space, for example where X lies on an algebraic curve or surface.

All the relevant steps (polynomial orthogonalization and computation of the Christoffel function, basic iteration, measure concentration) are carried out using standard MATLAB functions, such as the `qr` factorization and the `lsqnonneg` quadratic minimizer.

3. Illustrative Examples

In order to show the potentialities of the package, we present below a bivariate example on a nonconvex polygonal region with 27 sides, Ω say, resembling a flat and rough model of continental France; see Fig. 1. The region has been discretized by intersection with a 100×100 point grid of the minimal surrounding box, which in practice would correspond geographically to a discretization with stepsize of about 10 Km of the French territory. All the computations have been made in MATLAB R2017b on a 2.7 GHz Intel Core i5 CPU with 16GB RAM. The whole discretization mesh X of about 5700 points is concentrated at regression degree $n = 8$ into 153 sampling nodes and weights (a compression ratio around 38) keeping 95% G-efficiency ($\theta = 0.95$), in approximately 2 seconds.

In terms of deterministic regression error estimates, denoting by $L_n^{\mathbf{u}^*}$ the weighted least-squares operator corresponding to the Caratheodory-Tchakaloff concentration, $\mathbf{u}^*$, of the near-optimal design and by f a continuous function defined on the region, we can write

$$\max_{x \in X} \left| f(x) - L_n^{\mathbf{u}^*} f(x) \right| \leq \left(1 + \sqrt{N_n/\theta}\right) \min_{p \in \mathbb{P}_n^2} \max_{x \in X} |f(x) - p(x)|$$
$$\leq \left(1 + \sqrt{N_n/\theta}\right) \min_{p \in \mathbb{P}_n^2} \max_{x \in \Omega} |f(x) - p(x)| \,. \tag{5}$$

More precisely, in this example we get that the uniform regression error estimate on X (by sampling only at the Caratheodory-Tchakaloff concentrated support) is within a factor $1 + \sqrt{N_8/\theta} = 1 + \sqrt{45/0.95} \approx 7.88$ times the best uniform polynomial approximation of degree $n = 8$ to f on Ω (to be compared with a factor $1 + \sqrt{N_8} = 1 + \sqrt{45} \approx 7.71$ at full design optimality). If the resulting polynomial is not to one's satisfaction, one could always reconstruct the function f on the whole region from the grid values $\{L_n^{\mathbf{u}^*} f(x)\,,\ x \in X\}$ with a good accuracy (depending on smoothness), by any local or global interpolation scheme, such as splines or radial basis functions.

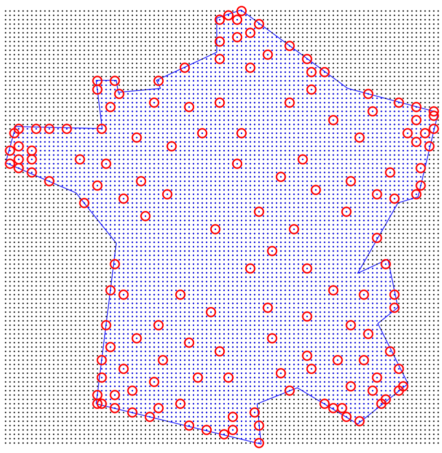

Figure 1: Caratheodory-Tchakaloff concentrated support (153 from 5746 points) for regression degree $n = 8$ on a nonconvex polygonal region after 27 iterations of Titterington's multiplicative algorithm (G-efficiency $\theta = 0.95$).

4. Impact

The computation of optimal designs for multivariate polynomial regression is a relevant issue in computational statistics and data analysis: cf., e.g., the classical textbook [1] and the recent paper [3], with the references therein. The approach proposed here is hybrid, in the sense that it starts by computing a design with a given threshold of G-optimality, say 95% to fix ideas, that could be more than appropriate in most applications, by performing only a few iterations of a basic multiplicative algorithm for design optimization (cf. [11, 12]).

At this level, the regressor quality is very good in the sense that the resulting approximation is nearly as good as it possibly can be relative to the best polynomial approximation (it should be noted that, of course, not all datasets can be well-fitted by polynomials). However, the cardinality of the support is typically still very high. Nevertheless, it is possible to strongly reduce the sampling cardinality, simply by resorting to recent implementations of Caratheodory-Tchakaloff discrete measure concentration: cf. [15, 16, 17, 18]. Only native MATLAB functions are involved in the computational process, namely `qr` factorizations of the relevant Vandermonde-like matrices and the `lsqnonneg` quadratic minimizer for the sparse nonnegative solution of the underlying moment system.

We are confident that the MATLAB package `CaTchDes`, in spite of its simplicity, will be useful in many applied contexts where bivariate and trivariate regression is a relevant tool, including, but not limited to, geo-spatial analysis.

5. Conflict of Interest

No conflict of interest exists: We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgements

Work partially supported by the Project BIRD181249 "NATIRESCO" and the DOR funds of the University of Padova, and by the GNCS-INdAM. This research has been accomplished within the RITA "Research ITalian network on Approximation".

References

[1] F. Pukelsheim, Optimal Design of Experiments, Classics in Applied Mathematics 50, SIAM, Philadelphia, 2006.

[2] T. Bloom, L. Bos, N. Levenberg, S. Waldron, On the convergence of optimal measures, Constr. Approx. 32 (2010).

[3] Y. De Castro, F. Gamboa, D. Henrion, R. Hess, J.-B. Lasserre, Approximate Optimal Designs for Multivariate Polynomial Regression, Ann. Statist. 47 (2019).

[4] V. Tchakaloff, Formules de cubatures mécaniques à coefficients non négatifs, (French) Bull. Sci. Math. 81 (1957).

[5] M. Putinar, A note on Tchakaloff's theorem, Proc. Amer. Math. Soc. 125 (1997).

[6] T. Bloom, L. Bos, N. Levenberg, The Asymptotics of Optimal Designs for Polynomial Regression, arXiv preprint: 1112.3735.

[7] J. Kiefer, J. Wolfowitz, The equivalence of two extremum problems, Canad. J. Math. 12 (1960).

[8] L. Bos, Some remarks on the Fejér problem for Lagrange interpolation in several variables, J. Approx. Theory 60 (1990).

[9] A. Mandal, W.K. Wong, Y. Yu, Algorithmic Searches for Optimal Designs, in: Handbook of Design and Analysis of Experiments, CRC, New York, 2015.

[10] L. Bos, F. Piazzon, M. Vianello, Near G-Optimal Tchakaloff Designs, submitted to Comput. Statistics.

[11] D.M. Titterington, Algorithms for computing d-optimal designs on a finite design space, in: Proc. 1976 Conference on Information Sciences and Systems, Baltimora, 1976.

[12] B. Torsney, R. Martin-Martin, Multiplicative algorithms for computing optimum designs, J. Stat. Plan. Infer. 139 (2009).

[13] C. Caratheodory, Über den Variabilittsbereich der Fourierschen Konstanten von positiven harmonischen Funktionen, Rend. Circ. Mat. Palermo 32 (1911).

[14] C.L. Lawson, R.J. Hanson, Solving Least-Squares Problems, revised reprint of the 1974 original, Classics in Applied Mathematics 15, SIAM, Philadelphia, 1995.

[15] C. Litterer, T. Lyons, High order recombination and an application to cubature on Wiener space, Ann. Appl. Probab. 22 (2012).

[16] F. Piazzon, A. Sommariva, M. Vianello, Caratheodory-Tchakaloff Subsampling, Dolomites Res. Notes Approx. DRNA 10 (2017).

[17] A. Sommariva, M. Vianello, Compression of multivariate discrete measures and applications, Numer. Funct. Anal. Optim. 36 (2015).

[18] M. Tchernychova, Caratheodory cubature measures, Ph.D. dissertation in Mathematics (supervisor: T. Lyons), University of Oxford, 2015.

[19] S. Foucart, H. Rahut, A Mathematical Introduction to Compressive Sensing, Birkhäuser, 2013.