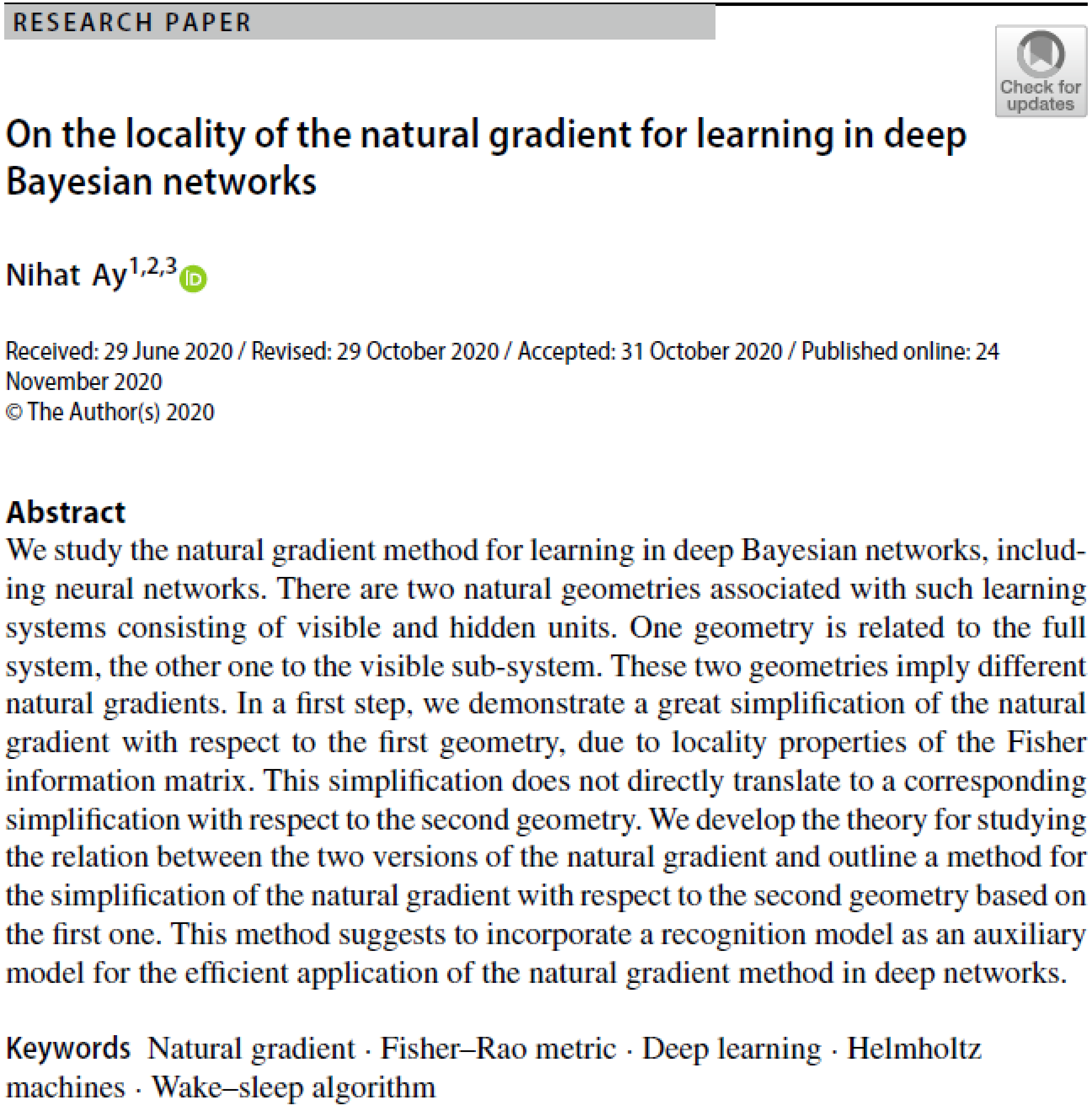


RESEARCH PAPER

On the locality of the natural gradient for learning in deep Bayesian networks

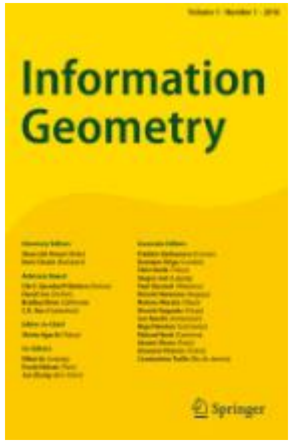
Nihat Ay[1,2,3]

Received: 29 June 2020 / Revised: 29 October 2020 / Accepted: 31 October 2020 / Published online: 24 November 2020

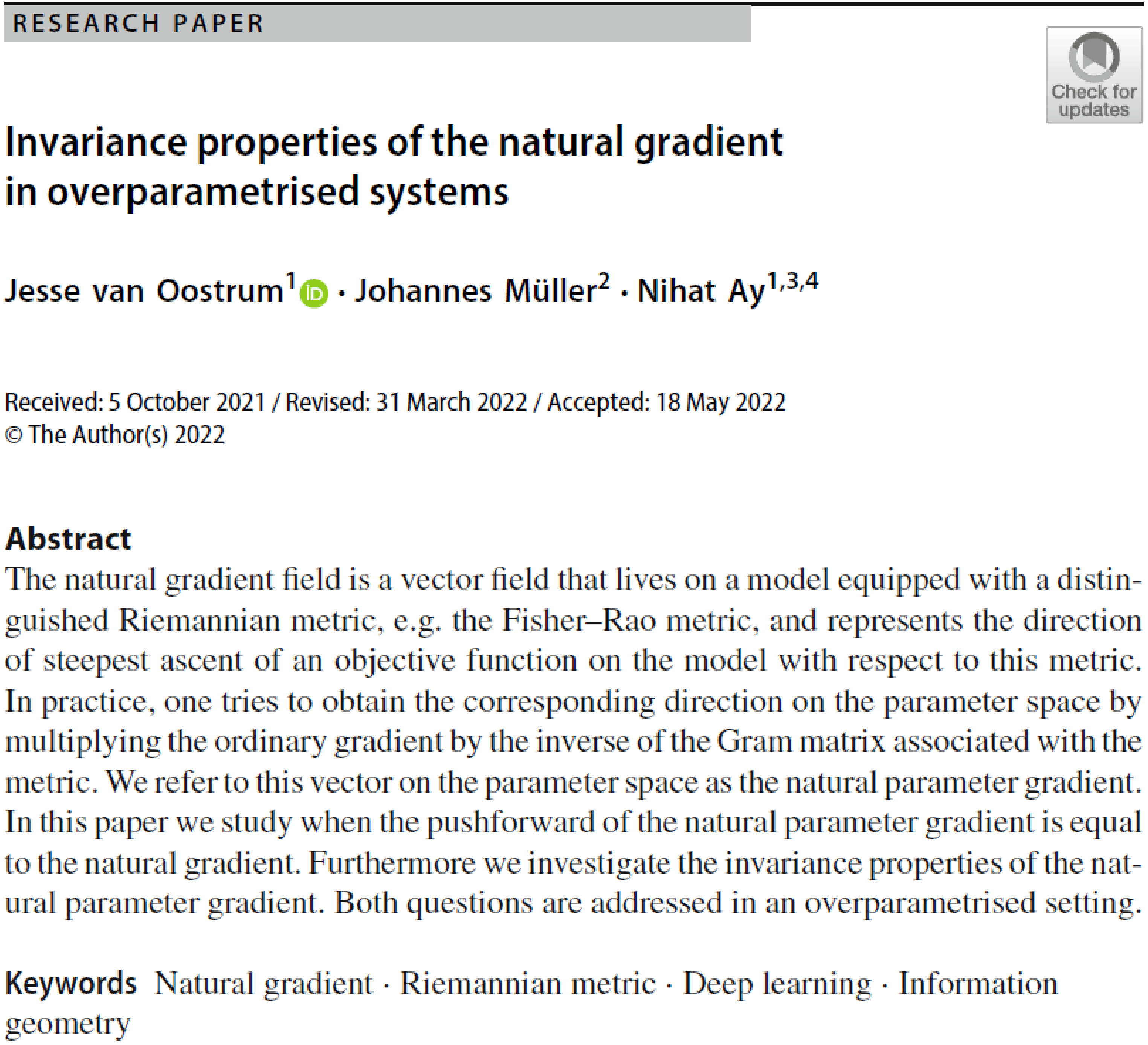
Abstract

We study the natural gradient method for learning in deep Bayesian networks, including neural networks. There are two natural geometries associated with such learning systems consisting of visible and hidden units. One geometry is related to the full system, the other one to the visible sub-system. These two geometries imply different natural gradients. In a first step, we demonstrate a great simplification of the natural gradient with respect to the first geometry, due to locality properties of the Fisher information matrix. This simplification does not directly translate to a corresponding simplification with respect to the second geometry. We develop the theory for studying the relation between the two versions of the natural gradient and outline a method for the simplification of the natural gradient with respect to the second geometry based on the first one. This method suggests to incorporate a recognition model as an auxiliary model for the efficient application of the natural gradient method in deep networks.

Keywords Natural gradient · Fisher–Rao metric · Deep learning · Helmholtz machines · Wake–sleep algorithm



Volume 6
Issue 1
Article 58



RESEARCH PAPER

Invariance properties of the natural gradient in overparametrised systems

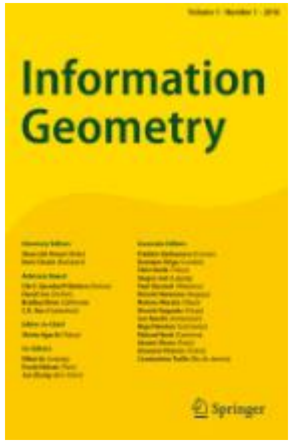
Jesse van Oostrum[1] · Johannes Müller[2] · Nihat Ay[1,3,4]

Received: 5 October 2021 / Revised: 31 March 2022 / Accepted: 18 May 2022

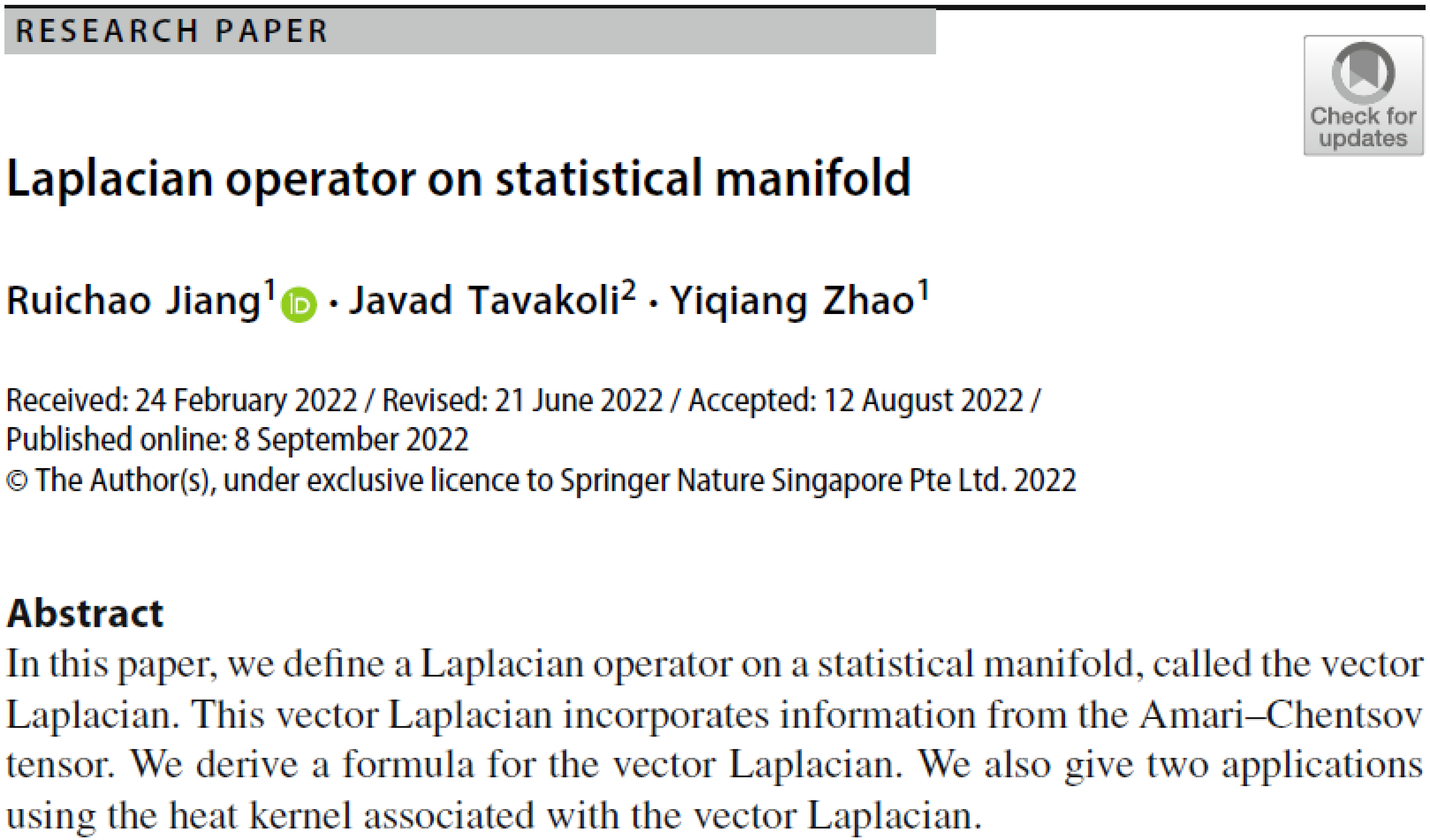
Abstract

The natural gradient field is a vector field that lives on a model equipped with a distinguished Riemannian metric, e.g. the Fisher–Rao metric, and represents the direction of steepest ascent of an objective function on the model with respect to this metric. In practice, one tries to obtain the corresponding direction on the parameter space by multiplying the ordinary gradient by the inverse of the Gram matrix associated with the metric. We refer to this vector on the parameter space as the natural parameter gradient. In this paper we study when the pushforward of the natural parameter gradient is equal to the natural gradient. Furthermore we investigate the invariance properties of the natural parameter gradient. Both questions are addressed in an overparametrised setting.

Keywords Natural gradient · Riemannian metric · Deep learning · Information geometry



Volume 6
Issue 1
Article 59



RESEARCH PAPER

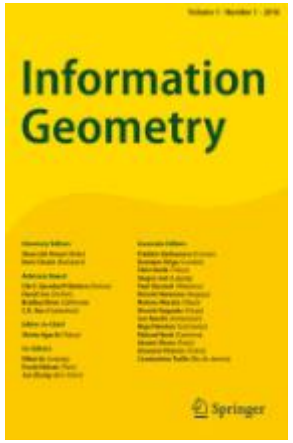
Laplacian operator on statistical manifold

Ruichao Jiang[1] · Javad Tavakoli[2] · Yiqiang Zhao[1]

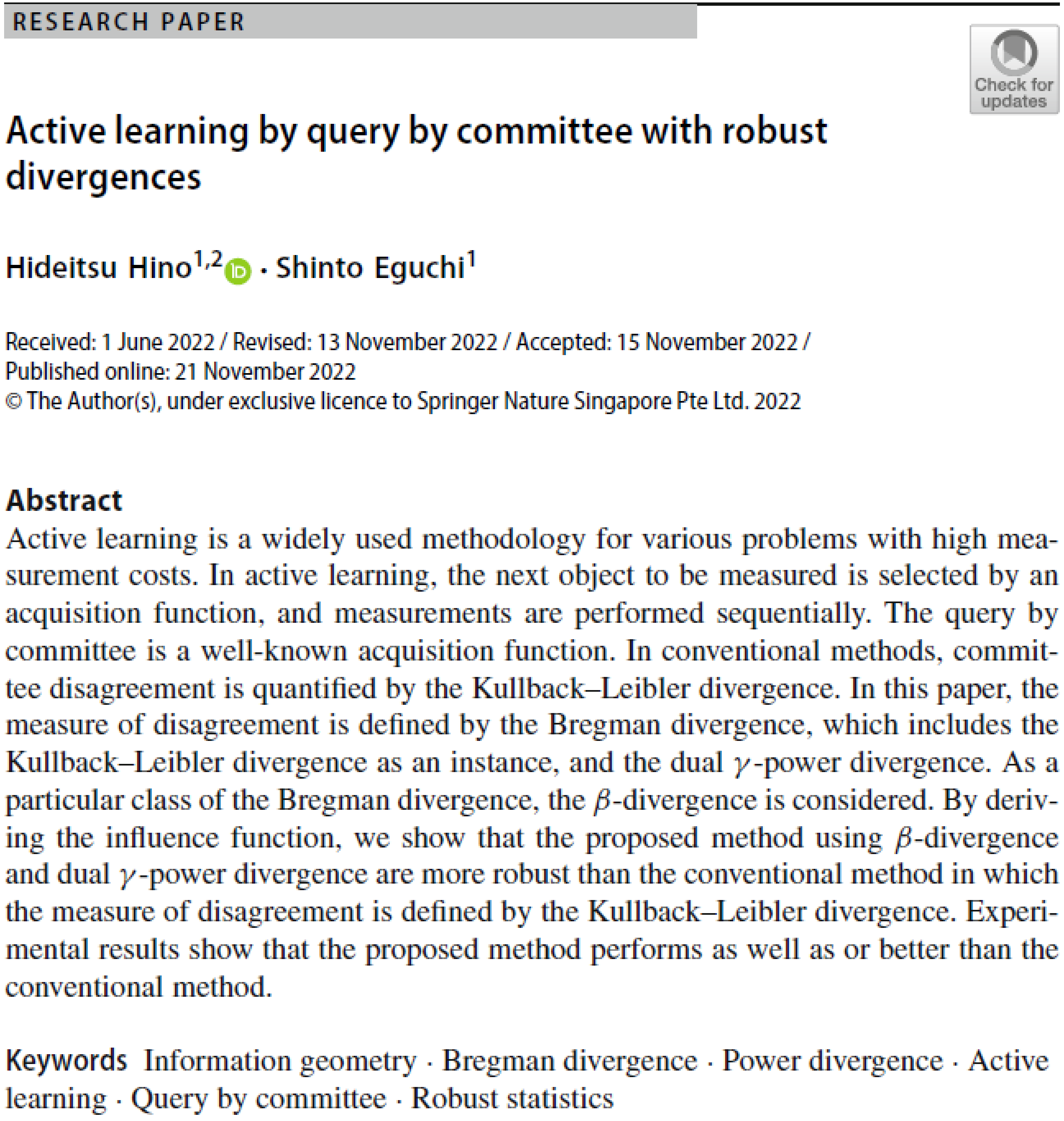
Received: 24 February 2022 / Revised: 21 June 2022 / Accepted: 12 August 2022 /
Published online: 8 September 2022

Abstract

In this paper, we define a Laplacian operator on a statistical manifold, called the vector Laplacian. This vector Laplacian incorporates information from the Amari–Chentsov tensor. We derive a formula for the vector Laplacian. We also give two applications using the heat kernel associated with the vector Laplacian.



Volume 6
Issue 1
Article 60



RESEARCH PAPER

Active learning by query by committee with robust divergences

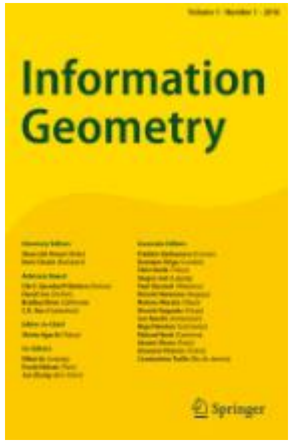
Hideitsu Hino[1,2] · Shinto Eguchi[1]

Received: 1 June 2022 / Revised: 13 November 2022 / Accepted: 15 November 2022 /
Published online: 21 November 2022

Abstract

Active learning is a widely used methodology for various problems with high measurement costs. In active learning, the next object to be measured is selected by an acquisition function, and measurements are performed sequentially. The query by committee is a well-known acquisition function. In conventional methods, committee disagreement is quantified by the Kullback–Leibler divergence. In this paper, the measure of disagreement is defined by the Bregman divergence, which includes the Kullback–Leibler divergence as an instance, and the dual γ-power divergence. As a particular class of the Bregman divergence, the β-divergence is considered. By deriving the influence function, we show that the proposed method using β-divergence and dual γ-power divergence are more robust than the conventional method in which the measure of disagreement is defined by the Kullback–Leibler divergence. Experimental results show that the proposed method performs as well as or better than the conventional method.

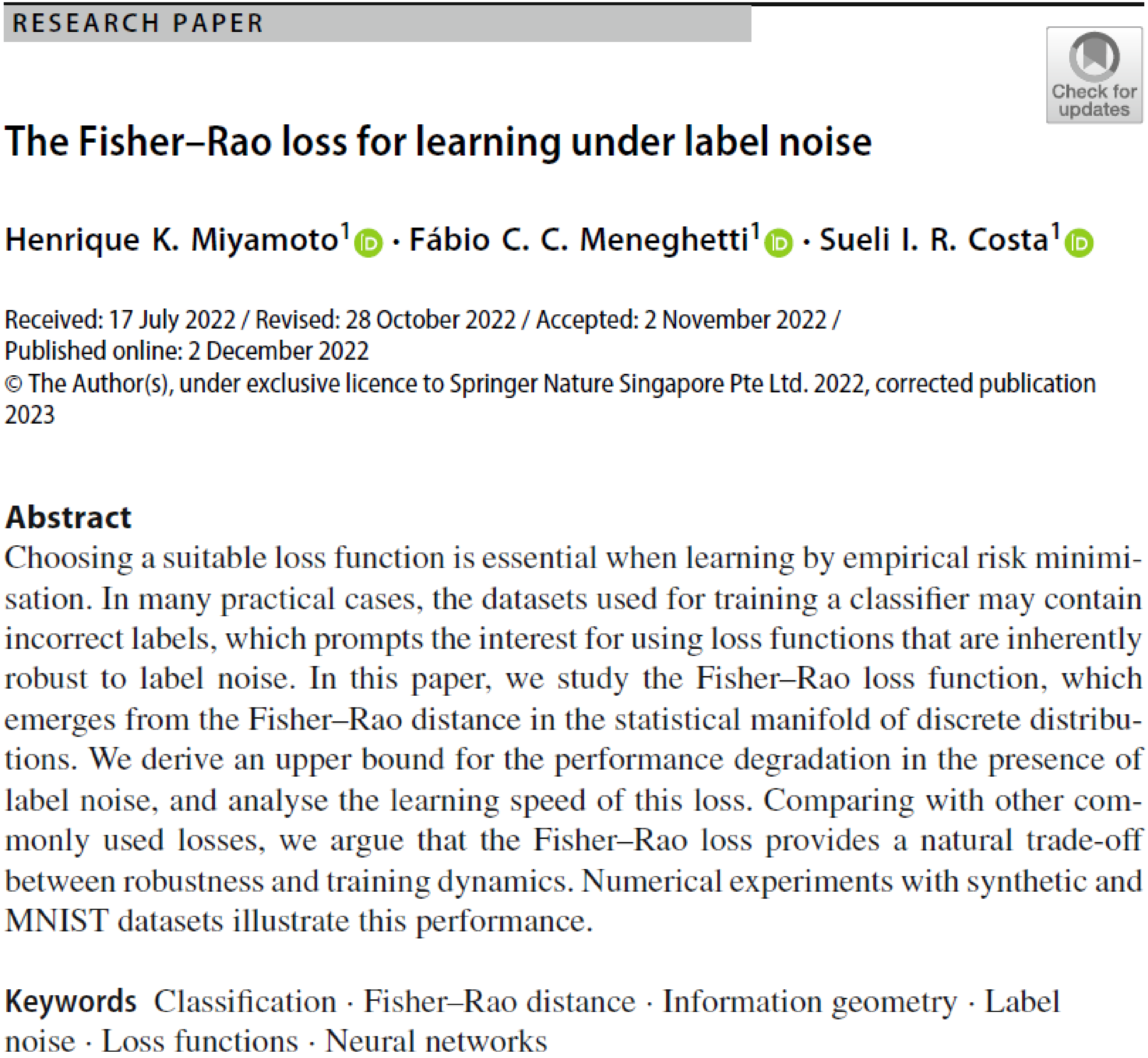
Keywords Information geometry · Bregman divergence · Power divergence · Active learning · Query by committee · Robust statistics



Information Geometry

Springer

Volume 6
Issue 1
Article 61



RESEARCH PAPER

The Fisher–Rao loss for learning under label noise

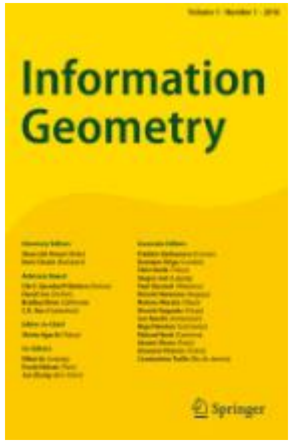
Henrique K. Miyamoto[1] · Fábio C. C. Meneghetti[1] · Sueli I. R. Costa[1]

Received: 17 July 2022 / Revised: 28 October 2022 / Accepted: 2 November 2022 /
Published online: 2 December 2022

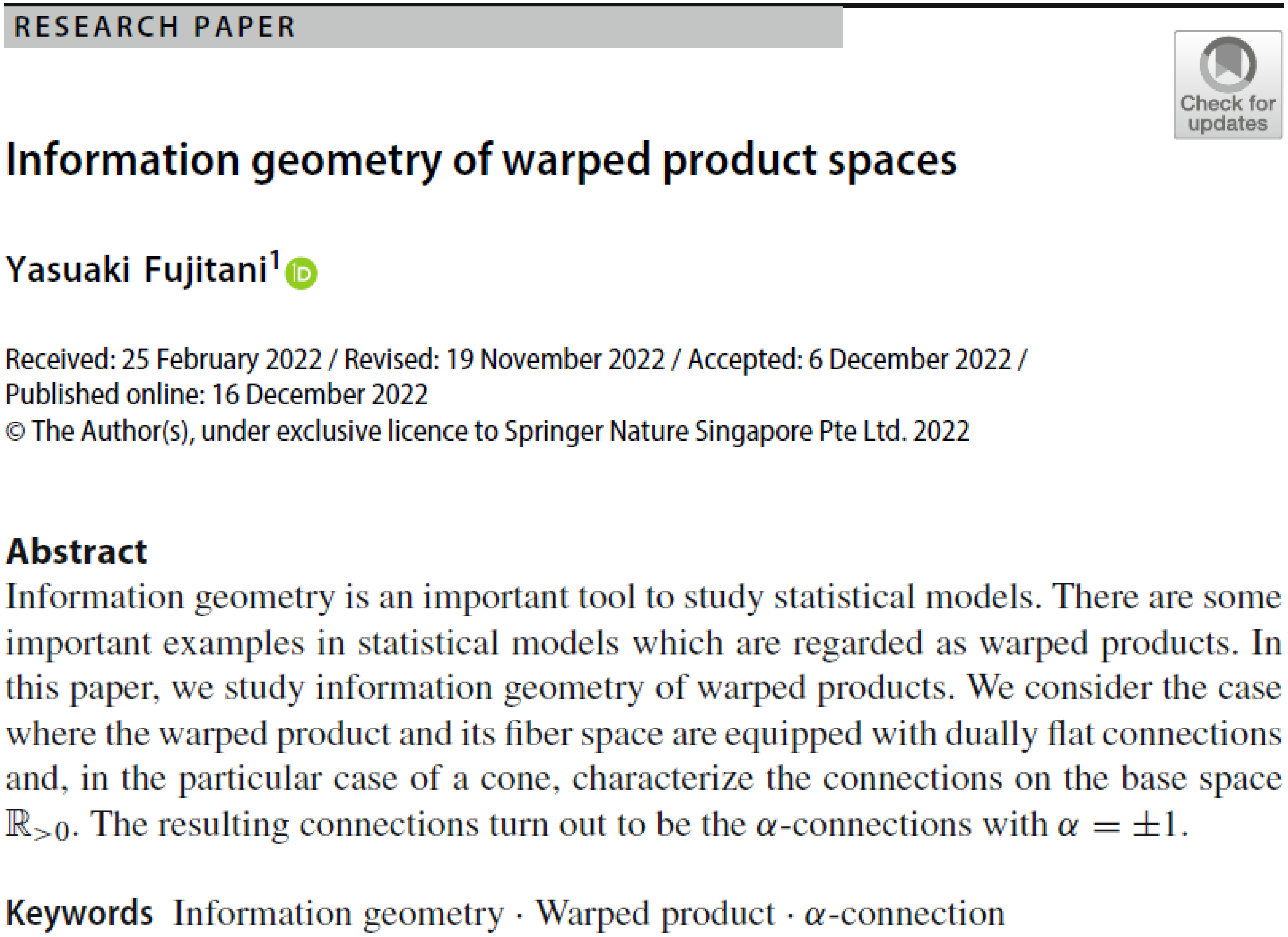
Abstract

Choosing a suitable loss function is essential when learning by empirical risk minimisation. In many practical cases, the datasets used for training a classifier may contain incorrect labels, which prompts the interest for using loss functions that are inherently robust to label noise. In this paper, we study the Fisher–Rao loss function, which emerges from the Fisher–Rao distance in the statistical manifold of discrete distributions. We derive an upper bound for the performance degradation in the presence of label noise, and analyse the learning speed of this loss. Comparing with other commonly used losses, we argue that the Fisher–Rao loss provides a natural trade-off between robustness and training dynamics. Numerical experiments with synthetic and MNIST datasets illustrate this performance.

Keywords Classification · Fisher–Rao distance · Information geometry · Label noise · Loss functions · Neural networks



Volume 6
Issue 1
Article 62



RESEARCH PAPER

Check for updates

Information geometry of warped product spaces

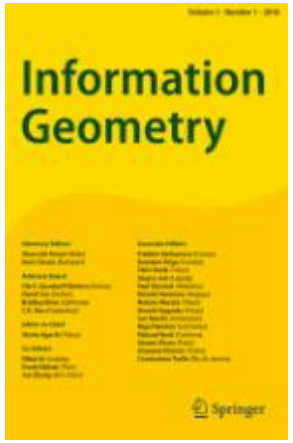
Yasuaki Fujitani[1]

Received: 25 February 2022 / Revised: 19 November 2022 / Accepted: 6 December 2022 /
Published online: 16 December 2022

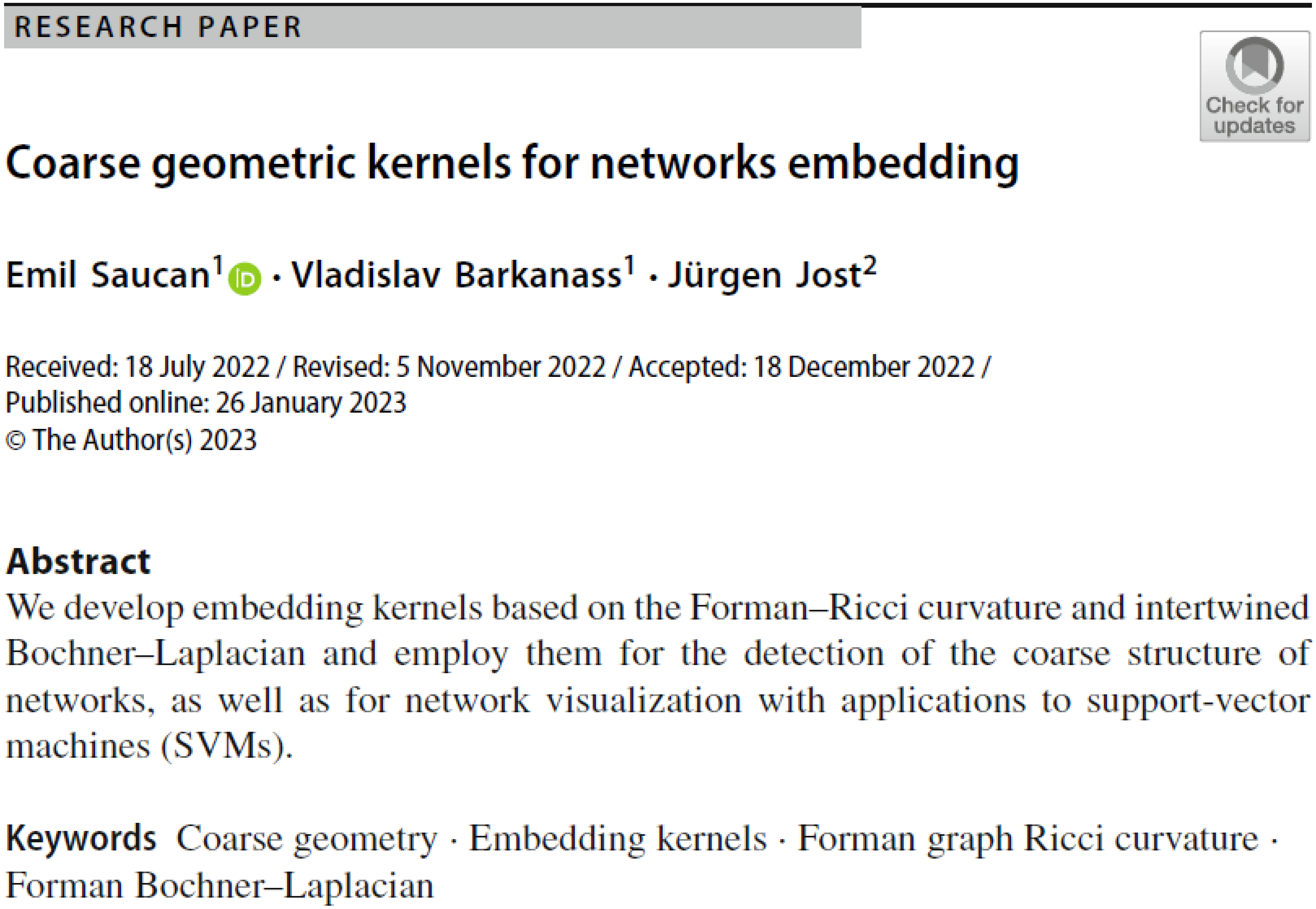
Abstract

Information geometry is an important tool to study statistical models. There are some important examples in statistical models which are regarded as warped products. In this paper, we study information geometry of warped products. We consider the case where the warped product and its fiber space are equipped with dually flat connections and, in the particular case of a cone, characterize the connections on the base space $\mathbb{R}_{>0}$. The resulting connections turn out to be the α-connections with $\alpha = \pm 1$.

Keywords Information geometry · Warped product · α-connection



Volume 6
Issue 1
Article 63



RESEARCH PAPER

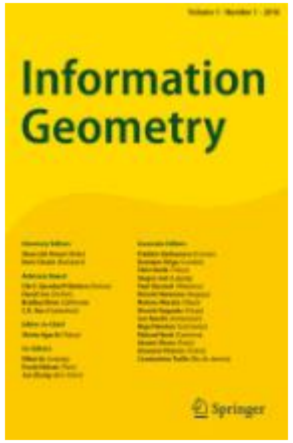
Coarse geometric kernels for networks embedding

Emil Saucan[1] · Vladislav Barkanass[1] · Jürgen Jost[2]

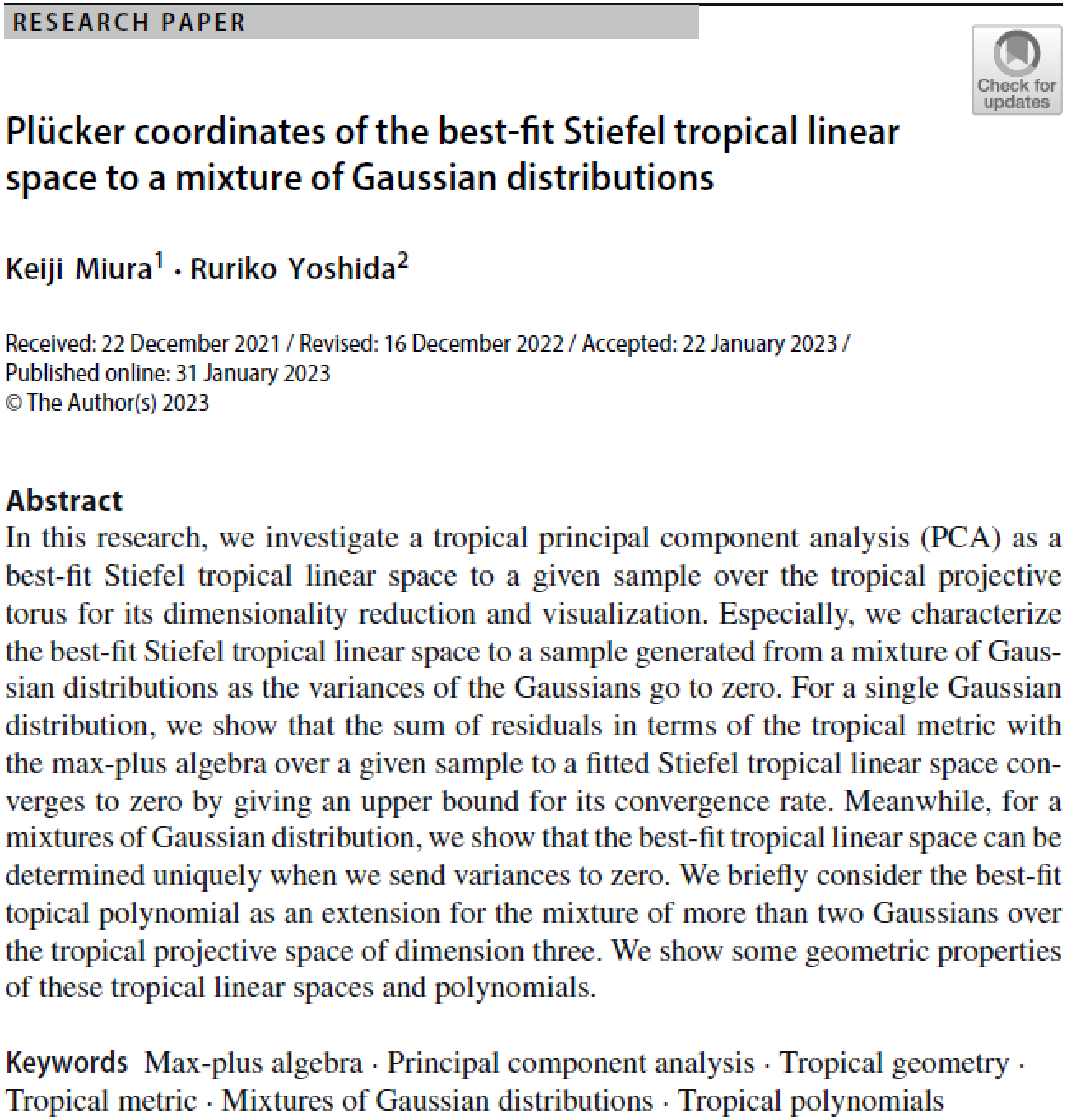
Received: 18 July 2022 / Revised: 5 November 2022 / Accepted: 18 December 2022 /
Published online: 26 January 2023

Abstract
We develop embedding kernels based on the Forman–Ricci curvature and intertwined Bochner–Laplacian and employ them for the detection of the coarse structure of networks, as well as for network visualization with applications to support-vector machines (SVMs).

Keywords Coarse geometry · Embedding kernels · Forman graph Ricci curvature · Forman Bochner–Laplacian



Volume 6
Issue 1
Article 64



RESEARCH PAPER

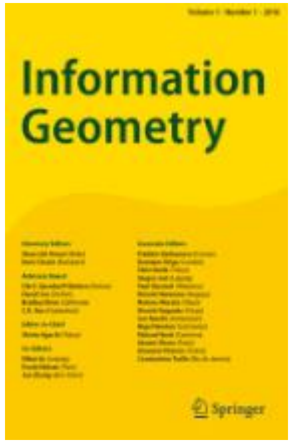
Plücker coordinates of the best-fit Stiefel tropical linear space to a mixture of Gaussian distributions

Keiji Miura[1] · Ruriko Yoshida[2]

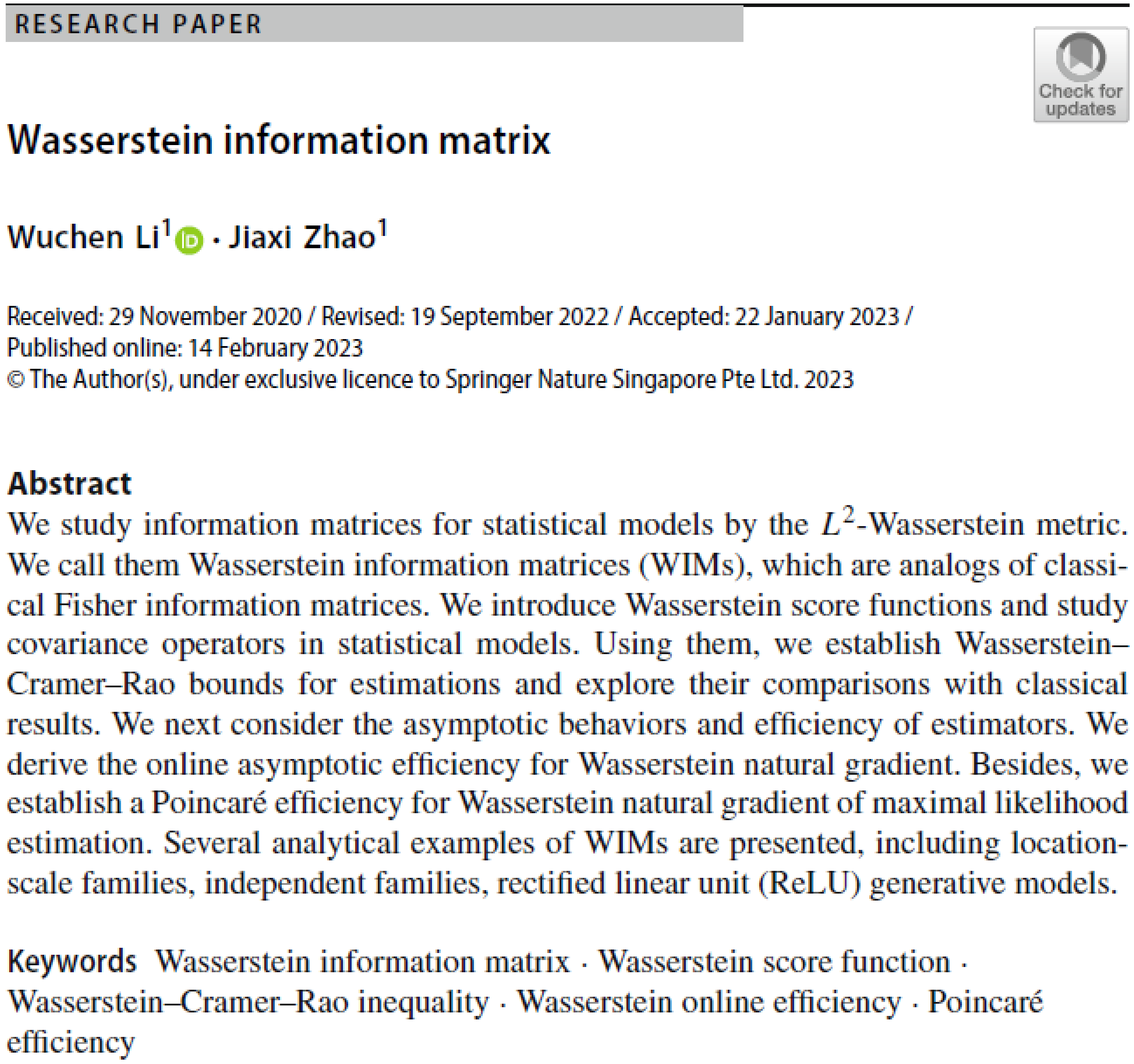
Received: 22 December 2021 / Revised: 16 December 2022 / Accepted: 22 January 2023 /
Published online: 31 January 2023

Abstract
In this research, we investigate a tropical principal component analysis (PCA) as a best-fit Stiefel tropical linear space to a given sample over the tropical projective torus for its dimensionality reduction and visualization. Especially, we characterize the best-fit Stiefel tropical linear space to a sample generated from a mixture of Gaussian distributions as the variances of the Gaussians go to zero. For a single Gaussian distribution, we show that the sum of residuals in terms of the tropical metric with the max-plus algebra over a given sample to a fitted Stiefel tropical linear space converges to zero by giving an upper bound for its convergence rate. Meanwhile, for a mixtures of Gaussian distribution, we show that the best-fit tropical linear space can be determined uniquely when we send variances to zero. We briefly consider the best-fit topical polynomial as an extension for the mixture of more than two Gaussians over the tropical projective space of dimension three. We show some geometric properties of these tropical linear spaces and polynomials.

Keywords Max-plus algebra · Principal component analysis · Tropical geometry · Tropical metric · Mixtures of Gaussian distributions · Tropical polynomials



Volume 6
Issue 1
Article 65



RESEARCH PAPER

Wasserstein information matrix

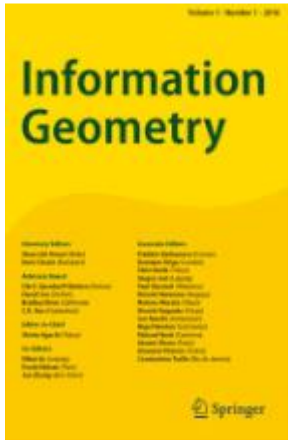
Wuchen Li[1] · Jiaxi Zhao[1]

Received: 29 November 2020 / Revised: 19 September 2022 / Accepted: 22 January 2023 /
Published online: 14 February 2023

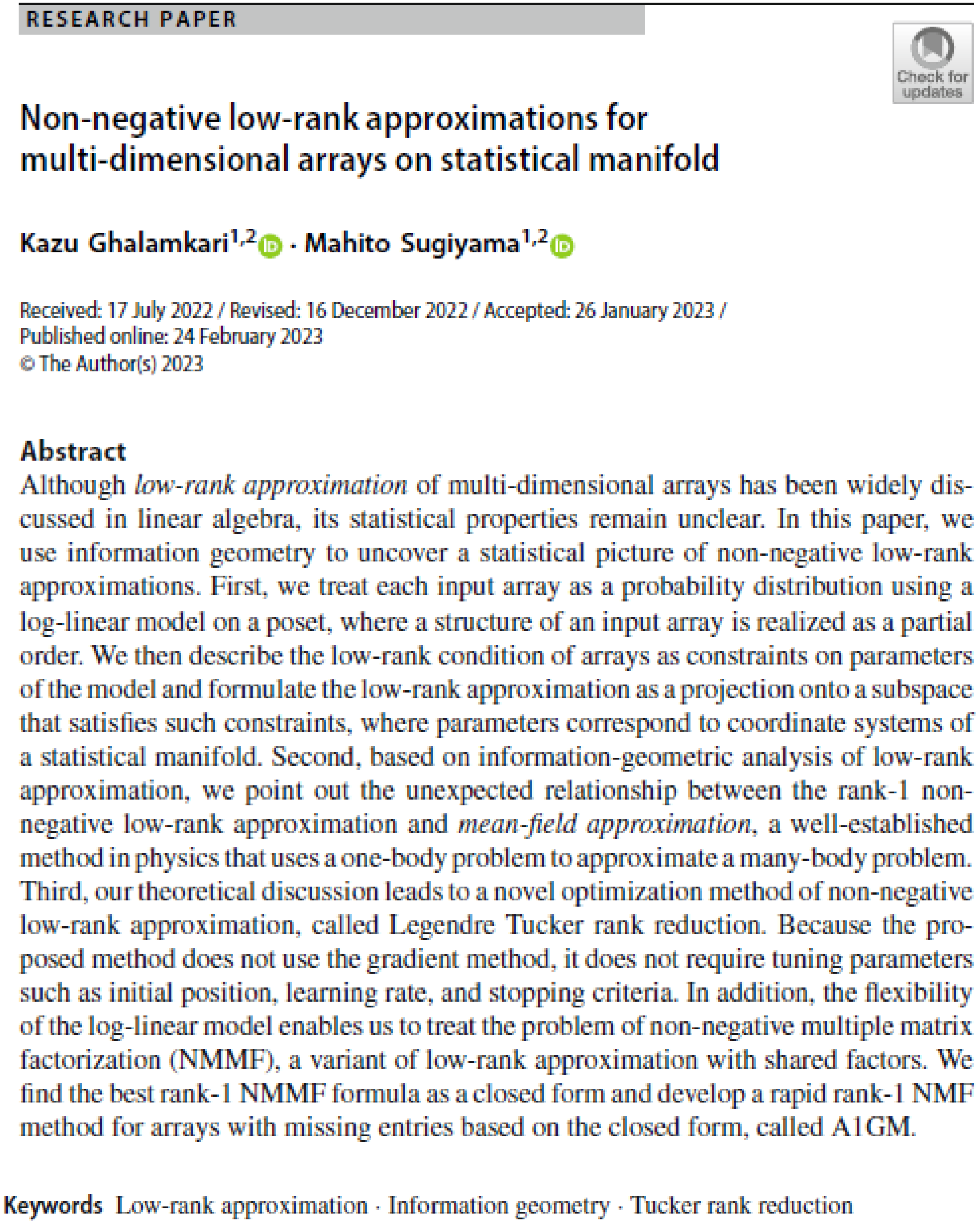
Abstract

We study information matrices for statistical models by the L^2-Wasserstein metric. We call them Wasserstein information matrices (WIMs), which are analogs of classical Fisher information matrices. We introduce Wasserstein score functions and study covariance operators in statistical models. Using them, we establish Wasserstein–Cramer–Rao bounds for estimations and explore their comparisons with classical results. We next consider the asymptotic behaviors and efficiency of estimators. We derive the online asymptotic efficiency for Wasserstein natural gradient. Besides, we establish a Poincaré efficiency for Wasserstein natural gradient of maximal likelihood estimation. Several analytical examples of WIMs are presented, including location-scale families, independent families, rectified linear unit (ReLU) generative models.

Keywords Wasserstein information matrix · Wasserstein score function · Wasserstein–Cramer–Rao inequality · Wasserstein online efficiency · Poincaré efficiency



Volume 6
Issue 1
Article 66



RESEARCH PAPER

Non-negative low-rank approximations for multi-dimensional arrays on statistical manifold

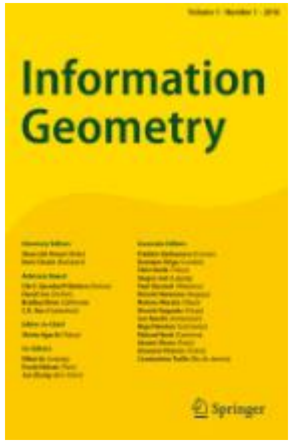
Kazu Ghalamkari[1,2] · Mahito Sugiyama[1,2]

Received: 17 July 2022 / Revised: 16 December 2022 / Accepted: 26 January 2023 /
Published online: 24 February 2023

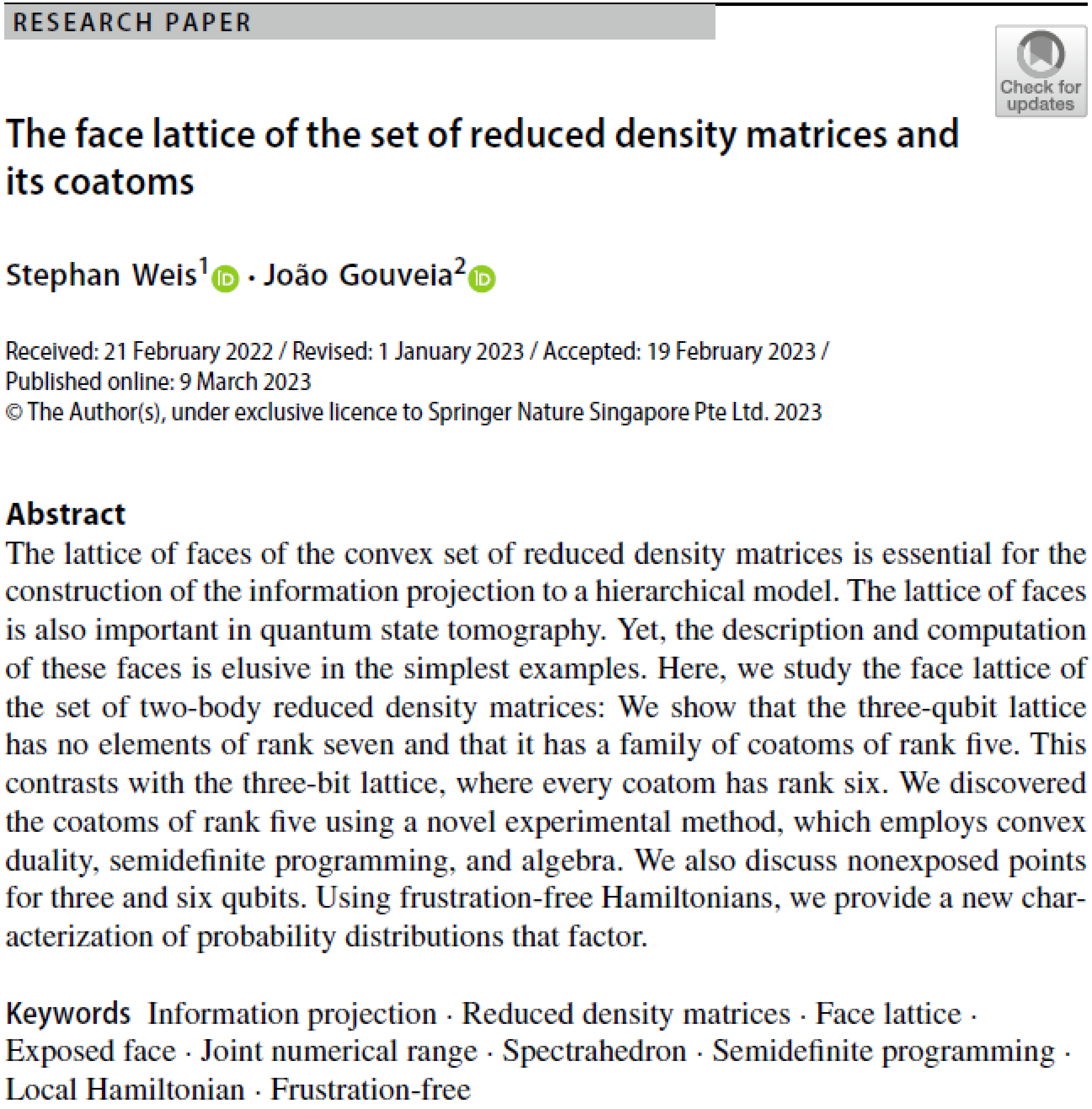
Abstract

Although *low-rank approximation* of multi-dimensional arrays has been widely discussed in linear algebra, its statistical properties remain unclear. In this paper, we use information geometry to uncover a statistical picture of non-negative low-rank approximations. First, we treat each input array as a probability distribution using a log-linear model on a poset, where a structure of an input array is realized as a partial order. We then describe the low-rank condition of arrays as constraints on parameters of the model and formulate the low-rank approximation as a projection onto a subspace that satisfies such constraints, where parameters correspond to coordinate systems of a statistical manifold. Second, based on information-geometric analysis of low-rank approximation, we point out the unexpected relationship between the rank-1 non-negative low-rank approximation and *mean-field approximation*, a well-established method in physics that uses a one-body problem to approximate a many-body problem. Third, our theoretical discussion leads to a novel optimization method of non-negative low-rank approximation, called Legendre Tucker rank reduction. Because the proposed method does not use the gradient method, it does not require tuning parameters such as initial position, learning rate, and stopping criteria. In addition, the flexibility of the log-linear model enables us to treat the problem of non-negative multiple matrix factorization (NMMF), a variant of low-rank approximation with shared factors. We find the best rank-1 NMMF formula as a closed form and develop a rapid rank-1 NMF method for arrays with missing entries based on the closed form, called A1GM.

Keywords Low-rank approximation · Information geometry · Tucker rank reduction



Volume 6
Issue 1
Article 67



RESEARCH PAPER

The face lattice of the set of reduced density matrices and its coatoms

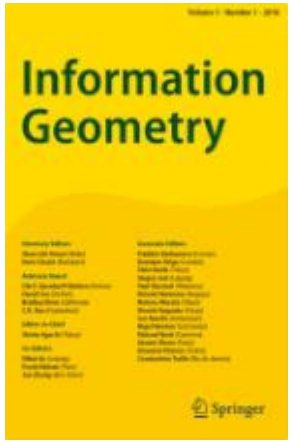
Stephan Weis[1] · João Gouveia[2]

Received: 21 February 2022 / Revised: 1 January 2023 / Accepted: 19 February 2023 /
Published online: 9 March 2023

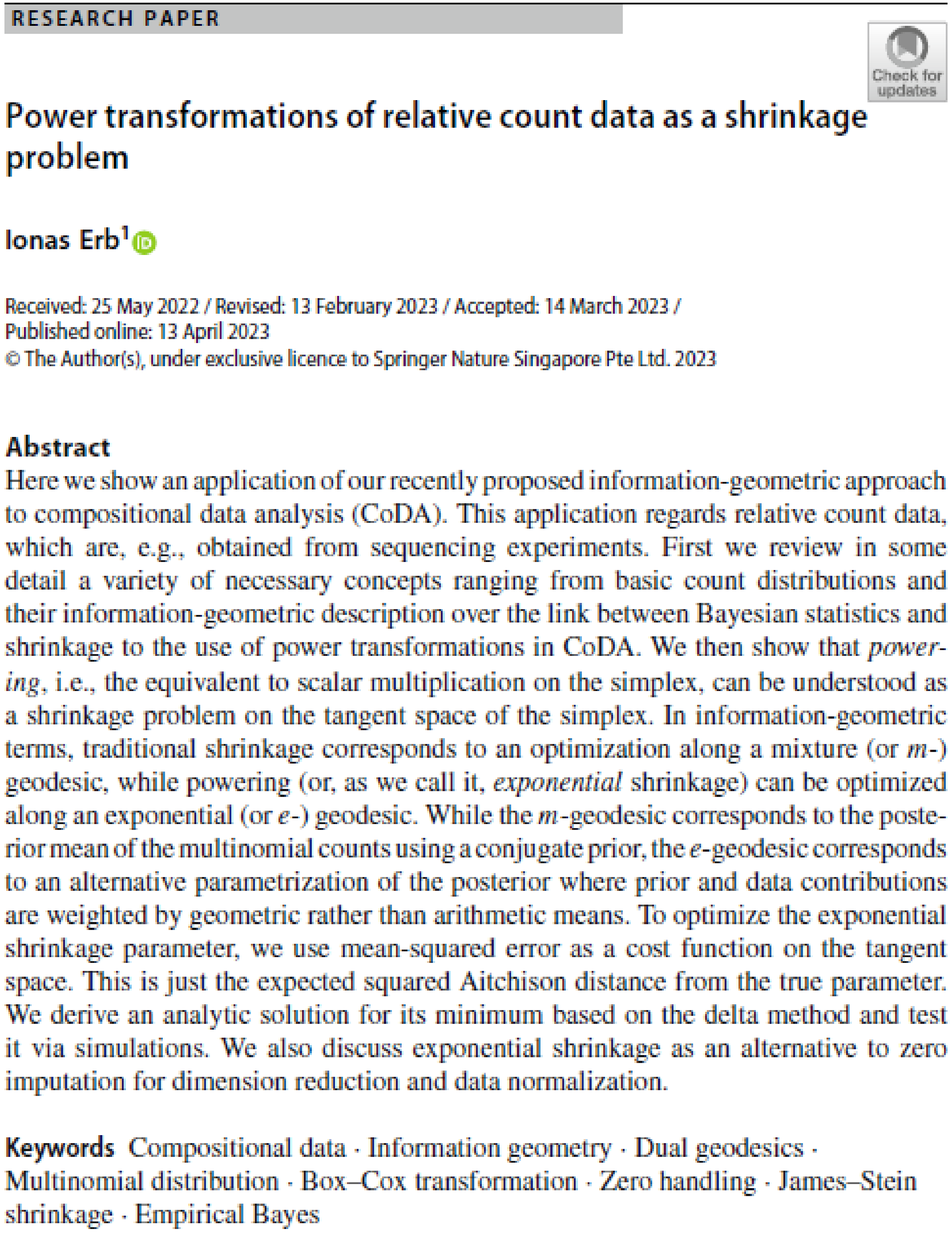
Abstract

The lattice of faces of the convex set of reduced density matrices is essential for the construction of the information projection to a hierarchical model. The lattice of faces is also important in quantum state tomography. Yet, the description and computation of these faces is elusive in the simplest examples. Here, we study the face lattice of the set of two-body reduced density matrices: We show that the three-qubit lattice has no elements of rank seven and that it has a family of coatoms of rank five. This contrasts with the three-bit lattice, where every coatom has rank six. We discovered the coatoms of rank five using a novel experimental method, which employs convex duality, semidefinite programming, and algebra. We also discuss nonexposed points for three and six qubits. Using frustration-free Hamiltonians, we provide a new characterization of probability distributions that factor.

Keywords Information projection · Reduced density matrices · Face lattice · Exposed face · Joint numerical range · Spectrahedron · Semidefinite programming · Local Hamiltonian · Frustration-free



Volume 6
Issue 1
Article 68



RESEARCH PAPER

Power transformations of relative count data as a shrinkage problem

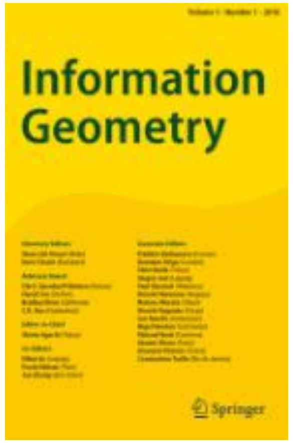
Ionas Erb[1]

Received: 25 May 2022 / Revised: 13 February 2023 / Accepted: 14 March 2023 /
Published online: 13 April 2023

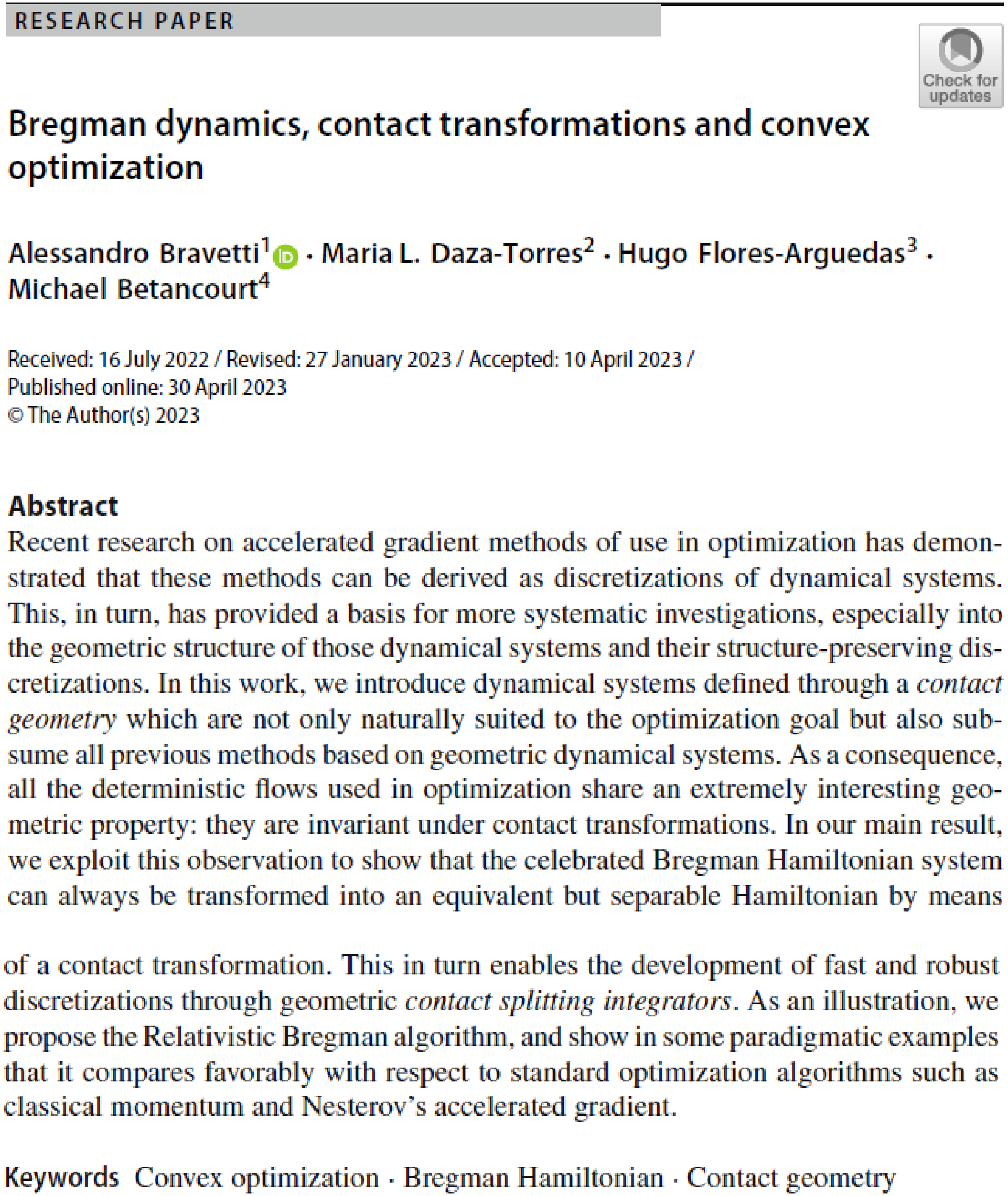
Abstract

Here we show an application of our recently proposed information-geometric approach to compositional data analysis (CoDA). This application regards relative count data, which are, e.g., obtained from sequencing experiments. First we review in some detail a variety of necessary concepts ranging from basic count distributions and their information-geometric description over the link between Bayesian statistics and shrinkage to the use of power transformations in CoDA. We then show that *powering*, i.e., the equivalent to scalar multiplication on the simplex, can be understood as a shrinkage problem on the tangent space of the simplex. In information-geometric terms, traditional shrinkage corresponds to an optimization along a mixture (or *m*-) geodesic, while powering (or, as we call it, *exponential* shrinkage) can be optimized along an exponential (or *e*-) geodesic. While the *m*-geodesic corresponds to the posterior mean of the multinomial counts using a conjugate prior, the *e*-geodesic corresponds to an alternative parametrization of the posterior where prior and data contributions are weighted by geometric rather than arithmetic means. To optimize the exponential shrinkage parameter, we use mean-squared error as a cost function on the tangent space. This is just the expected squared Aitchison distance from the true parameter. We derive an analytic solution for its minimum based on the delta method and test it via simulations. We also discuss exponential shrinkage as an alternative to zero imputation for dimension reduction and data normalization.

Keywords Compositional data · Information geometry · Dual geodesics · Multinomial distribution · Box–Cox transformation · Zero handling · James–Stein shrinkage · Empirical Bayes



Volume 6
Issue 1
Article 69



RESEARCH PAPER

Bregman dynamics, contact transformations and convex optimization

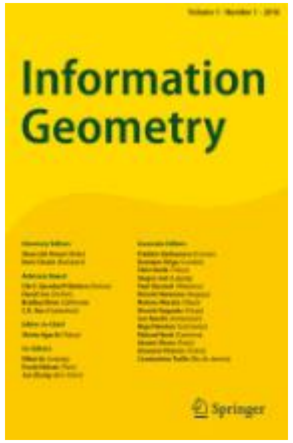
Alessandro Bravetti[1] · Maria L. Daza-Torres[2] · Hugo Flores-Arguedas[3] · Michael Betancourt[4]

Received: 16 July 2022 / Revised: 27 January 2023 / Accepted: 10 April 2023 /
Published online: 30 April 2023

Abstract

Recent research on accelerated gradient methods of use in optimization has demonstrated that these methods can be derived as discretizations of dynamical systems. This, in turn, has provided a basis for more systematic investigations, especially into the geometric structure of those dynamical systems and their structure-preserving discretizations. In this work, we introduce dynamical systems defined through a *contact geometry* which are not only naturally suited to the optimization goal but also subsume all previous methods based on geometric dynamical systems. As a consequence, all the deterministic flows used in optimization share an extremely interesting geometric property: they are invariant under contact transformations. In our main result, we exploit this observation to show that the celebrated Bregman Hamiltonian system can always be transformed into an equivalent but separable Hamiltonian by means of a contact transformation. This in turn enables the development of fast and robust discretizations through geometric *contact splitting integrators*. As an illustration, we propose the Relativistic Bregman algorithm, and show in some paradigmatic examples that it compares favorably with respect to standard optimization algorithms such as classical momentum and Nesterov's accelerated gradient.

Keywords Convex optimization · Bregman Hamiltonian · Contact geometry



Volume 6
Issue 1
Article 70