SIAG/OPT Views and News

A Forum for the SIAM Activity Group on Optimization

Volume 25 Number 1 January 2017

Contents

Comments from the Editors

We thank the contributors to the 35th issue of the SIAM Activity Group on Optimization's newsletter. The theme of this issue is "optimization and imaging", with an article on phase retrieval by D. Russell Luke and nonnegative matrix factorization by Nicolas Gillis. We hope you enjoy this latest installment of Views and News.

Congratulations to the new officers for the SIAM Activity Group on Optimization: Tamás Terlaky, Andreas Waechter, Michael Friedlander, and James Luedtke. Thank you for agreeing to serve our activity group for the next three years.

Many of you have written to opt for an electronic copy of Views and News; for the others among you, please do not hesitate to contact us to opt out of receiving physical copies.

As always, we welcome your feedback, (e-)mailed directly to us or to siagoptnews@lists.mcs.anl.gov. Suggestions for new issues, comments, and papers are always welcome!

Stefan Wild, Editor
Mathematics and Computer Science Division, Argonne National Laboratory, USA, wild@anl.gov, http://www.mcs.anl.gov/~wild
Jennifer Erway, Editor
Department of Mathematics, Wake Forest University, USA, erwayjb@wfu.edu, http://www.wfu.edu/~erwayjb

Articles

Phase Retrieval, What's New?

D. Russell Luke
Institut für Numerische und Angewandte Mathematik, Universität Göttingen, 37083 Göttingen, Germany.
E-Mail: r.luke@math.uni-goettingen.de
http://num.math.uni-goettingen.de/~r.luke

Ask an engineer to solve a problem and she will come back in a day or so with something that seems to work well enough most of the time. Ask a mathematician to solve the same problem and he will return many months later with an exact but unimplementable solution to a different problem. I'm sure most readers of this newsletter have heard some variation of that joke. But a true story lies somewhere in there, a story that is writ large with the phase retrieval problem.

The phase retrieval problem has been around for more than a century, and it is solved tens of thousands of times each second, mostly by physicists. Phase retrieval plays a central role in the x-ray imaging experiments conducted by researchers here in Göttingen, where we are in the last 5-year funding cycle of a 15-year collaborative research center studying nanoscale photonic imaging (Deutsche Forschungsgemeinschaft CRC755). The center consists of experimental physicists and biomolecular physicists building new instruments and observation techniques (one of those techniques, STED, won center participant Stefan Hell a Nobel Prize in 2014) as well as mathematicians studying algorithms, image processing, and statistics. Phase retrieval is an applied mathematician's dream problem: it is central to many imaging modalities, it is simple to state, numerical routines for its solution abound, and it is mathematically interesting in ways that solving systems of linear equations will never be.

Nick Trefethan wrote in his introduction to a 2002 SIAM Review article on phase retrieval that I wrote together with Jim Burke and Rick Lyon [42], "A Google search of 'phase retrieval' returns 271,000 records." Almost fifteen years later, a Google search returns 364,000 records (with the safe search on). A Web of Science^{TM} database search back to 1945 yields 8,924 results, 7,189 of those since 2002, more than half of those since 2011. A lot of new interest has been expressed recently in particular within some corners of the statistics and applied mathematics communities. Apparently, money

also is at stake, as I learned in preparation for this article. So, if the problem has been around for so long and people are already solving it–Hauptman and Karle won a Nobel Prize in 1985 for solving the crystallographic phase retrieval problem–what's all the recent fuss about?

For those who don't know what phase retrieval is, it is simply stated as follows: Find $x \in C \subset \mathbb{C}^n$ such that $|(F_k x)|_j = b_{jk}$, where for $k = 1, 2, \ldots, K$ the mapping $F_k : \mathbb{C}^n \to \mathbb{C}^m$ is linear and $b_{jk} \in \mathbb{R}_+$ for all $j = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, K$. The classical problem comes from diffraction imaging where the set C is some a priori constraint like support or nonnegativity, and the mapping F_k is a Fourier transform of some kind. This includes the Fresnel transform and some defocused or otherwise imperfect Fraunhofer transform. The structure of the model is simple: the components of the vectors $F_k x$ must lie on circles of a given radius in the complex plane. Unfortunately, circles are not convex (the sphere does not contain any line segment joining any two points on the sphere), causing all sorts of problems, both mathematical and practical. The first statement of the phase problem that I could find goes back to a letter from J. W. Strutt (Lord Rayleigh) to Michaelson in 1892 [53]. Lord Rayleigh was pessimistic about the prospects of breaking the phase barrier unless a priori information about symmetry was known. A solution to the band-limited phase problem is a zero of a related complex polynomial. I'll come back to the qualifier "band limited" in a moment, but ignore that detail for now. In the 1950s Akutowicz [1, 2] showed that the one-dimensional phase retrieval problem without any a priori constraints has many solutions, since by the fundamental theorem of algebra, all 1D complex polynomials factor into products of monomials. A lot of workarounds for the unconstrained 1D phase retrieval problem have been developed since the 1970s [17, 25, 31, 50, 52, 57, 58, 7], all of which involve adding a constraint implicitly or explicitly. Some recent progress on the 1D problem has come from initialization techniques that land one in a neighborhood of the minimum phase solution [30] where the usual nonlinear programming techniques can perform reliably.

At the end of the 1970s Bruck and Sodin [10] pointed out that the fundamental barrier to unconstrained 1D phase retrieval does not apply in higher dimensions since, magically, polynomials of dimension two or more *almost never* factor. This conformed nicely with the unreasonable success of the simple Gerchberg-Saxton [23] and HIO [21] algorithms for 2D phase retrieval proposed a few years earlier. Shortly thereafter Hayes [24] proved that for band-limited signals, the 2D phase retrieval problem has unique solutions, almost surely, up to rotations, shifts, and reflections. One might conclude that the book on phase retrieval was closed a long time ago, except that the theory didn't quite match up with practice as nicely as one would hope. The first hint that the story is more complicated came from the algorithms themselves. They worked fairly well a lot of the time, but one of the more popular approaches, HIO, *never* worked in the usual sense of convergence to a fixed point. Still today, people continue to apply HIO according to the following recipe: run 10-40 iterations of HIO; then apply several passes of Gerchberg-Saxton to clean up the image; publish.

I have the fortune of being coauthor with Heinz Bauschke and Patrick Combettes of a paper on phase retrieval algorithms that gets a steady stream of citations [3]. Unfortunately, fewer people read it than cite it. We started from the premise that really only a handful of good first-order algorithms exist and that anything that works is probably a tweak of one of those. It was known before our paper that Gerchberg-Saxton and the *error reduction algorithm* [21] are simply alternating projections (one of the handful of good algorithms). We were able partially to identify HIO for the case of a support constraint alone by showing the correspondence between this procedure and the now ubiquitous Douglas-Rachford algorithm. In a follow-up paper [4] we showed that HIO with a support and nonnegativity constraint becomes a *different* fixed-point iteration, what we called the hybrid projection reflection (HPR) method (*not* one of the handful). This fundamental change in the fixed-point mapping by a seemingly minor change in the constraint structure is not obvious when the algorithms are written in the format favored in the optics literature. At the same time the HPR method was presented, Veit Elser introduced his *difference map* [19], which for certain parameter values coincides with the Douglas-Rachford and HPR algorithms, again depending on the constraint structure [38]. The instabilities of these algorithms together with the insight provided by the more mathematical prescription of the algorithms led me to propose a relaxation of the Douglas-Rachford algorithm, which I called RAAR, that has fixed points when Douglas-Rachford does not [39]. At that time, alternating projections, Douglas-Rachford, HPR, and RAAR algorithms were understood only for convex problems. In the convex setting Douglas-Rachford can be identified with the alternating directions method of multipliers [22], which is currently popular for large-scale problems. For nonconvex problems, however, our understanding of Douglas-Rachford and even alternating projections, and hence everything else close to these, pretty much evaporated. Since then a lot of quiet, patient work has been done in the variational analysis community to develop the theory of first-order methods for nonconvex problems, and much of the missing theory behind the success of these algorithms for phase retrieval is in place. But I get ahead of myself.

Almost any article on phase retrieval in the applied mathematics literature will start with a statement like "The phase retrieval problem is found in many different areas of science and engineering, such as x-ray crystallography, astronomy, diffraction imaging, and more." So I contacted several physicists and astronomers to find out from them what is new in phase retrieval. One place where efficient solutions to the phase retrieval problem is of vital importance is the W. M. Keck Observatory. The Keck instruments need to correct for random aberrations in the Earth's atmosphere in order to compete with instruments such as the Hubble Space Telescope. The shape of the atmosphere is encoded in the phase of the observations. Sam Ragland, an adap-

tive optics scientist at Keck, told me that their instruments use Shack-Hartmann sensors to measure the phase directly. This is a hardware solution to the phase problem. Keck's wavefront controllers operate at a rate of 2 kHz, 1 kHz in practice. Ragland did say that they were testing a computational phase-diversity algorithm (phase retrieval with several defocused images), for which they are at the moment using just the Gerchberg-Saxton algorithm.

The cost of the Shack-Hartmann sensors is photons, which are in short supply in more modern x-ray imaging set-ups. The dominant approach here is computational phase retrieval, although recent proposals involve adding random masks to the imaging systems in order to avoid nonuniqueness in the numerical reconstructions. With regard to algorithms for phase retrieval, Elser, an expert in diffraction x-ray imaging at Cornell University, has seen "nothing significant" in the past 5 years. He is even more dismissive of the impact of random masks: "The integrity of those phase masks has to be established at the same resolution as their intended application – and for that you need a phasing system that works at that resolution!" This is a fundamental challenge for nanoscale (i.e., x-ray) imaging. Pierre Thibault, an expert in blind ptychography (the second worst phase retrieval problem one can encounter) at the University of Southampton, is more circumspect. Algorithms such as alternating projections, HPR, Douglas-Rachford, RAAR, and the difference map work just fine, according to Thibault; "the biggest bottleneck is hardware." While Thibault has found no use for random masks in his work, he would not discount the possibility that the idea of randomly generated data could have unforeseen applications. One such possibility he mentioned was Fourier ptychographic microscopy, which iteratively stitches together a number of variably (i.e., randomly) illuminated, low-resolution intensity images in Fourier space to produce a wide-field, high-resolution complex sample image [56, 59]. The idea that more Fourier measurements can improve the performance of phase retrieval algorithms has been around for a while. For crystallographic phase retrieval there is only so much information you can get out of the intensity measurements. But in the early 2000s it was recognized that for noncrystallographic measurements one is not limited to sampling on a fixed lattice and that oversampling dramatically improves numerical reconstructions [45]. Unfortunately, this improved performance is attributed to some form of uniqueness. This is curious since a few moments of reflection on elementary Fourier analysis is all that is needed to be convinced that oversampling has nothing to do with uniqueness. Increased, but still finite, sampling in the Fourier domain just pushes the error created by trying to reconstruct a compactly supported object from finitely many Fourier measurements to some level below either numerical or experimental precision.

The oversampling justification is just one example of a fixation on uniqueness that has overshadowed the most obvious structural problem for phase retrieval: existence. Remember the "band-limited" qualifier in Hayes' result cited earlier. What that means is a compactly supported Fourier transform. And, what that means is that the object must be periodic. At this point many people retreat to the discrete Fourier transform, which is a unitary linear operator, about the best kind of operator one can have – except that the best physical model we have for describing what we measure in any physical experiment is a sample of the continuous Fourier transform. And when you implement phase retrieval on a computer, you cannot avoid implicitly setting the values of the part of the Fourier spectrum that you do not measure to zero. So almost any constraint–in particular compact support–that you place on the object whose Fourier transform you sample will be inconsistent with the measurement. One exception is crystallography, where (perfect) crystals are indeed periodic and the Fourier transform can be assumed reasonably to be band limited. The most exciting and difficult imaging challenges today, however, are in noncrystallographic "single-shot" x-ray imaging [48, 20]. I mentioned that blind ptychography is the second hardest type of phase retrieval problem you might encounter. Blind ptychography [26, 55, 44, 29] is akin to reading a fragment of an ancient text in a script you have never seen, with a pair of glasses borrowed from someone you have never met, and being asked to reconstruct simultaneously the script and the prescription of the glasses. For single-shot x-ray imaging the script consists of 3D figurines floating randomly in midair and of which you get only brief flashes from a strobe light. Phase retrieval is the easy part for reconstructions from single-shot data - the challenging part is the tomographic reconstruction of the Fourier data. Unfortunately, for much of the more recent mathematical work directed at phase retrieval to have any traction, uniqueness is essential; but for these more modern applications even existence of a solution to the model equations together with qualitative constraints cannot be taken for granted.

The recent work in applied mathematics on phase retrieval has its roots in a series of now-famous papers by Candès and Tao [14, 13], which showed that under certain conditions on the matrix generating an affine subspace of $\mathbb{R}^n$ (called the *restricted isometry property* in the literature), there is a unique sparsest point in the subspace and, moreover, this point is the point with smallest ℓ_1-norm. When the space is a space of matrices, the uniqueness is up to orbits, and the elements of the orbits have smallest nuclear norm. This has sparked a wave of papers on convex (and even nonconvex) relaxation in the signal processing community since finding a point in an affine subspace with minimum norm is a convex optimization problem while the problem of minimizing the counting function subject to an affine constraint is nonconvex and NP-hard [47]. Bloomensath and Davies [8, 9] ran against the current, however, and examined a simple forward-backward prox-algorithm, iterative hard thresholding, for solving a slightly different problem of minimizing the norm of the residual in the image space subject to a sparsity inequality constraint in the domain. They showed that under an asymmetric generalization of the restricted isometry conditions required for the correspondence of the nonconvex sparsity optimization problem and its convex re-

laxation, iterative hard thresholding converges globally to the global solution of the sparsity-constrained residual minimization problem. Their work inspired me and my students Robert Hesse and Patrick Neumann to show that the same, or similar, conditions also guarantee global linear convergence of alternating projections for the problem of finding the intersection of an affine subspace and the set of vectors with sparsity less than a given value [28]. We also showed that the asymmetric restricted isometry conditions for global convergence of alternating projections imply *transversality* of the range of the transpose of the matrix generating the affine subspace and the orthogonal complement of all subspaces of dimension twice the dimension of the sparsest element. A consequence, which has not been much explored, is that there cannot be any locally optimal solutions for this nonconvex problem other than the unique global minimum. Apparently, convex relaxations are not even needed for problems with this structure.

The connection to phase retrieval, pointed out first by Candés, Eldar, Strohmer, and Voroninski [12], uses a well-known trick from conic programming for turning a quadratic function on $\mathbb{R}^n$ into a linear function on the space of $n \times n$ matrices. The price to pay for this is in going from a problem with n unknowns to a problem with n^2 unknowns. Structurally, however, the problem in the space of matrices is the same as the spasity optimization problem above. There was some hope that this would lead to a breakthrough in phase retrieval by solving a convex problem in the space of matrices. But the prospect of squaring the number of unknowns should have given pause to even the most ardent booster of Moore's law. Phase lift, as this idea is called in the literature, has not proven to be a reasonable computational strategy. In the past two years there has been some backing away from phase lift as an algorithm, and more direct nonconvex methods are again being proposed (see [32] and references therein). No one from the applications side that I spoke to for this article was aware of these newer methods, however. One reason could be that none of the newer methods has been compared with the methods favored in the optics community. Reference to methods such as Gerchberg-Saxton and HIO and the identification with classical algorithms made in [3] seems to be obligatory in recent articles, but they appeal to missing theory behind these methods in to order avoid direct comparisons.

I'm happy to report that we actually now know quite a bit about the classical algorithms for the phase retrieval problem. I mentioned in the beginning that only a handful of good first-order algorithms exist. Among these are steepest descent (which includes averaged projections, the Misell algorithm for phase diversity, and many of the schemes proposed in the past few years), forward-backward prox schemes (which include projected gradients, hard- and soft-thresholding, and accelerations), backward-backward prox schemes (which include alternating projections and hence Gerchberg-Saxton), and Douglas-Rachford (to which category I assign HIO, RAAR, and the difference map). The standard "old" algorithms for phase retrieval are all based on projectors that are composed and averaged in some fashion. Implicit in this is a *feasibility* formulation of the phase problem, that is, to find some point in the intersection of the set of points satisfying the constraints implied by the data measurements and the set of points satisfying qualitative constraints such as support and nonnegativity. This is an extremely powerful modeling approach since it is easy to introduce new constraints without changing one's algorithmic approach. It also lays bare the success and failure of various methods.

When one settles on a feasibility model for a problem, uniqueness is almost irrelevant. The two important cases for feasibility are consistent and inconsistent. In the consistent case the sets have at least one point in common; in the inconsistent case, the sets have no point in common. Most of the progress on the nonconvex convergence theory for the good algorithms above has been for the consistent case. Based on a series of papers exposing the structure of alternating projections in increasingly inhospitable environments [15, 35, 34, 40, 5, 18, 49, 27], we know now that alternating projections applied to consistent phase retrieval problems is locally linearly convergent at points of intersection except in the unlikely case that the constraints are tangential. The nonconvex theory of the Douglas-Rachford algorithm for consistent problems is also fairly well understood in settings that cover phase retrieval [27, 51].

As I argued above, however, the real-world phase retrieval problem is hopelessly inconsistent. This inconsistency can be observed by simply running the Douglas-Rachford algorithm on your favorite experimental data set. You will observe the iterates moving around seemingly chaotically, sometimes looking like something well structured before wandering off to nonsense. This behavior is a consequence of the fact that the Douglas-Rachford mapping applied to inconsistent feasibility has no fixed points. For convex problems this is not a serious issue since the *shadow sequence* of the iterates can be shown to converge [6]. For nonconvex problems, however, all bets are off, and this explains the instability of the HIO algorithm for phase retrieval.

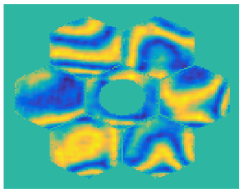


Figure 1: Representative iterate of a noisy JWST test wavefront recovered with the Douglas-Rachford algorithm. For a movie showing instability of the algorithm, go to http://num.math.uni-goettingen.de/proxtoolbox

The RAAR algorithm [38, 39] is a relaxation of the Douglas-Rachford algorithm that is guaranteed to have fixed points for a strong enough relaxation. One can easily verify [29, 40] that the regularity of the RAAR mapping for feasibility-based phase retrieval satisfies one of two condi-

tions that together are sufficient to guarantee local linear convergence developed in [51], namely, that the constraint sets are *superregular*. What remains to be shown in order to guarantee that the RAAR algorithm is locally linearly convergent for phase retrieval is that the RAAR fixed-point mapping is *metrically regular* [33] (in some appropriate sense) at fixed points. Abstract formulas for showing metric regularity exist and rely on computing the coderivative of the fixed-point operator [16, 46], but executing these calculations and verifying the conditions for phase retrieval is complicated.

Another way to understand the RAAR algorithm outlined in [39] is as the Douglas-Rachford algorithm applied not to the problem of minimizing the sum of two indicator functions but rather to the problem of minimizing the weighted sum of the squared distance to one of the sets plus the indicator of the other set. Loock and Plonka [36, 37] took this idea further to apply the RAAR algorithm not to the squared distance to a set but rather to the ℓ_1-norm of the wavelet transform of the object to be recovered. They were able to show that the iterates of the RAAR algorithm in this setting are at least bounded. I am optimistic that the remaining open issues concerning the convergence of the RAAR algorithm for phase retrieval will soon be resolved. As pointed out in [39], the power of Douglas-Rachford and its relaxations in the context of feasibility is that it can be tuned to have far fewer fixed points (i.e., locally optimal points), than algorithms such as alternating projections or steepest descents.

For the alternating projections algorithm applied to phase retrieval, the picture is fairly complete. The results are local, as one can expect from any nonconvex problem. For practical, inconsistent phase retrieval, recent work with Matt Tam and Thao Nguyen shows that alternating projections must converge locally linearly to local best approximation points except in rare degenerate cases [43, Theorem 5.10 and Example 5.16]. This result allows for *error bounds* as stopping criteria for this algorithm. What we cannot say is whether the fixed points of the algorithm are good, and this has been the main point of criticism. But let us return to the observation that under conditions similar to, albeit stronger than, those used to justify convex relaxations in sparsity optimization, alternating projections converges globally linearly to a unique global solution in that setting. We can then conjecture that alternating projections for phase retrieval with enough measurements converges globally linearly to a globally optimal best approximation point. A nice opportunity exists here for the two strands of analysis that have been picking away at the phase retrieval problem–one from the variational analysis side and the other from sparsity optimization–to merge productively. A strength of the theory sparked by [14, 13] is that it can say something about how much information is needed before one can reasonably expect nice things to happen on a global scale, and this has nothing to do with convexity or the quantitative local analysis.

The algorithms and phenomena discussed here can be explored in the ProxToolbox [41], which is a slowly growing collection of demonstrations of simple first-order methods built on prox-operators. Following the example of Buckheit and Donoho [11] and more recent calls for reproducible research [54], we are trying to make available all numerical demonstrations that have supported our publications. This effort will expand to data and algorithms from the broader Nanoscale Photonic Imaging Collaborative Research Center at Göttingen. Phase Focus Limited of Sheffield, UK, claims intellectual property rights on iterative routines for ptychography and has sued researchers. No academic researcher has the means to challenge such assertions and this has put a chill on efforts to disseminate information, but it does not appear to be an outright barrier. In this age of increased suspicion of science and the scientific method, it is all the more important to make our work as transparent and accessible as possible.

Acknowledgments. This work was supported in part by German Israeli Foundation Grant G-1253-304.6 and Deutsche Forschungsgemeinschaft Collaborative Research Center SFB755.

REFERENCES

[1] E. J. Akutowicz. On the determination of the phase of a Fourier integral, I. *Trans. Amer. Math. Soc.*, 83:179–192, 1956.

[2] E. J. Akutowicz. On the determination of the phase of a Fourier integral, II. *Proc. Amer. Math. Soc.*, 8:234–238, 1957.

[3] H. H. Bauschke, P. L. Combettes, and D. R. Luke. Phase retrieval, error reduction algorithm and Fienup variants: a view from convex feasibility. *J. Opt. Soc. Amer. A.*, 19(7):1334–45, 2002.

[4] H. H. Bauschke, P. L. Combettes, and D. R. Luke. Finding best approximation pairs relative to two closed convex sets in Hilbert spaces. *J. Approx. Theory*, 127:178–92, 2004.

[5] H. H. Bauschke, D. R. Luke, H. M. Phan, and X. Wang. Restricted normal cones and the method of alternating projections: applications. *Set-Valued Var. Anal.*, 21:475–501, 2013.

[6] H.H. Bauschke and W.M. Moursi. On the Douglas-Rachford algorithm. arXiv:1604.04603, 2016.

[7] R. Beinert and G. Plonka. Ambiguities in one-dimensional phase retrieval of structured functions. *Proc. Appl. Math. Mech.*, 15(1):653–654, 2015.

[8] T. Blumensath and M. Davies. Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. A.*, 27(3):265–274, 2009.

[9] T. Blumensath and M. Davies. Normalised iterative hard thresholding; guaranteed stability and performance. *IEEE J. Sel. Topics Signal Process.*, 4(2):298–309, 2010.

[10] Y. M. Bruck and L. G. Sodin. On the ambiguity of the image reconstruction problem. *Opt. Comm.*, 30(3):304–308, Sept. 1979.

[11] J. Buckheit and D. Donoho. Wavelab and reproducible research. In A. Antoniadis et al., editors, *Wavelets and Statistics*. Springer-Verlag, New York, 1995.

[12] E. Candès, Y. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM J. Imaging Sci.*, 6(1):199–225, 2013.

[13] E. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203–4215, December 2005.

[14] E. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theory*, 52(12):5406–5425, December 2006.

[15] P. L. Combettes and H. J. Trussell. Method of successive projections for finding a common point of sets in metric spaces. *J. Opt. Theory and Appl.*, 67(3):487–507, 1990.

[16] A. L. Dontchev and R. T. Rockafellar. *Implicit Functions and Solution Mapppings*. Springer-Verlag, New York, second edition, 2014.

[17] A. J. J. Drenth, A. Huiser, and H. Ferwerda. The problem of phase retrieval in light and electron microscopy of strong objects. *Opt. Acta*, 22:615–628, 1975.

[18] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. Transversality and alternating projections for nonconvex sets. *Found. Comput. Math.*, 15(6):1637–1651, 2015.

[19] V. Elser. Phase retrieval by iterated projections. *J. Opt. Soc. Amer. A*, 20(1):40–55, 2003.

[20] V. Elser. Strategies for processing diffraction data from randomly oriented particles. *Ultramicroscopy*, 111:788–792, 2011.

[21] J. R. Fienup. Reconstruction of an object from the modulus of its Fourier transform. *Opt. Lett.*, 3(1):27–29, 1978.

[22] D. Gabay. Applications of the method of multipliers to variational inequalities. In *Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems*, pages 299–331. North-Holland, Amsterdam, 1983.

[23] R. W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

[24] M. H. Hayes. The unique reconstruction of multidimensional sequences from Fourier transform magnitude or phase. In H. Stark, editor, *Image Recovery: Theory and Application*. Academic Press, New York, 1987.

[25] M. H. Hayes and A. V. Oppenheim. Signal reconstruction from phase or magnitude. *IEEE Trans. Acoustics, Speech, and Signal Process.*, ASSP-28(6):672–680, December 1980.

[26] R. Hegerl and W. Hoppe. Dynamische theorie der kristallstrukturanalyse durch elektronenbeugung im inhomogenen primrstrahlwellenfeld. *Ber. Bunsenges. Phys. Chem*, 74(11):1148–1154, 1970.

[27] R. Hesse and D. R. Luke. Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. *SIAM J. Optim.*, 23(4):2397–2419, 2013.

[28] R. Hesse, D. R. Luke, and P. Neumann. Alternating projections and Douglas–Rachford for sparse affine feasibility. *IEEE Trans. Signal. Process.*, 62(18):4868–4881, 2014.

[29] R. Hesse, D. R. Luke, S. Sabach, and M. Tam. The proximal heterogeneous block implicit-explicit method and application to blind ptychographic imaging. *SIAM J. Imaging Sci.*, 8(1):426–457, 2015.

[30] K. Huang, Y. C. Eldar, and N. D. Sidiropoulos. Phase retrieval from 1D Fourier measurements: convexity, uniqueness, and algorithms. *IEEE Trans. Signal Process.*, 64(23):6105–6117, 2016.

[31] A. Huiser and H. Ferwerda. The problem of phase retrieval in light and electron microscopy of strong objects, II: on the uniqueness and stability of object reconstruction procedures using two defocused images. *Optica Acta*, 23:445–456, 1976.

[32] K. Jaganathan, Y. C. Eldar, and B. Hassibi. Phase retrieval: an overview of recent developments. arXiv:1510.07713, 2015.

[33] A. Y. Kruger, D. R. Luke, and Nguyen H. T. Set regularities and feasibility problems. *Math. Programming B*, 2016. arXiv:1602.04935.

[34] A. S. Lewis, D. R. Luke, and J. Malick. Local linear convergence of alternating and averaged projections. *Found. Comput. Math.*, 9(4):485–513, 2009.

[35] A. S. Lewis and J. Malick. Alternating projections on manifolds. *Math. Oper. Res.*, 33:216–234, 2008.

[36] S. Loock and G. Plonka. Phase retrieval for Fresnel measurements using a shearlet sparsity constraint. *Inverse Probl.*, 30, 055005, 2014.

[37] S. Loock and G. Plonka. Iterative phase retrieval with sparsity constraints. *PAMM*, 16(1):835–836, 2016.

[38] D. R. Luke. Relaxed averaged alternating reflections for diffraction imaging. *Inverse Probl.*, 21:37–50, 2005.

[39] D. R. Luke. Finding best approximation pairs relative to a convex and a prox-regular set in Hilbert space. *SIAM J. Optim.*, 19(2):714–739, 2008.

[40] D. R. Luke. Local linear convergence of approximate projections onto regularized sets. *Nonlinear Anal.*, 75:1531–1546, 2012.

[41] D. R. Luke. ProxToolbox, 2016. http://num.math.uni-goettingen.de/proxtoolbox/.

[42] D. R. Luke, J. V. Burke, and R. G. Lyon. Optical wavefront reconstruction: theory and numerical methods. *SIAM Rev.*, 44:169–224, 2002.

[43] D. R. Luke, Nguyen H. Thao, and M. T. Tam. Quantitative convergence analysis of iterated expansive, set-valued mappings. http://arxiv.org/abs/1605.05725, May 2016.

[44] A. M. Maiden and J. M. Rodenburg. An improved ptychographical phase retrieval algorithm for diffractiveimaging. *Ultramicroscopy*, 109:1256–1262, 2009.

[45] J. Miao, P. Charalambous, J. Kirz, and D. Sayre. Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, 400:342–344, 1999.

[46] B.S. Mordukhovich. *Variational Analysis and Generalized Differentiation, I: Basic Theory; II: Applications*. Grundlehren Math. Wiss. Springer-Verlag, New York, 2006.

[47] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24:227–234, 1995.

[48] R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, and J. Hajdu. Potential for biomolecular imaging with femtosecond x-ray pulses. *Nature*, 406:752–757, 2000.

[49] D. Noll and A. Rondepierre. On local convergence of the method of alternating projections. *Found. Comput. Math.*, 16(2):425–455, 2016.

[50] A.V. Oppenheim and R.W. Schafer. *Digital Signal Processing*. Prentice Hall, Englewood Cliffs, NJ, 1975.

[51] H. Phan. Linear convergence of the Douglas–Rachford method for two closed sets. *Optimization*, 65:369–385, 2016.

[52] T. Quatieri and A.V. Oppenheim. Iterative techniques for minimum phase signal reconstruction from phase or magnitude. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-29(6):1187–1193, Dec. 1981.

[53] J. W. Strutt (Lord Rayleigh). On the interference bands of approximately homogeneous light; in a letter to Prof. A. Michelson. *Phil. Mag.*, 34:407–411, 1892.

[54] V. Stodden, M. McNutt, D. H. Bailey, E. Deelman, Y. Gil, B. Hanson, M. A. Heroux, J. P. A. Ioannidis, and M. Taufer. Enhancing reproducibility for computational methods. *Science*, 354(6317):1240–1241, 2016.

[55] P. Thibault, M. Dierolf, O. Bunk, A. Menzel, and F. Pfeiffer. Probe retrieval in ptychographic coherent diffractive imaging. *Ultramicroscopy*, 109:338–343, 2009.

[56] L. Tian, X. Li, K. Ramchandran, and L. Waller. Multiplexed coded illumination for Fourier ptychography with an LED array microscope. *Biomedical Optics Express*, 5(7):2376–2389, 2014.

[57] P. van Toorn and H. Ferwerda. The problem of phase retrieval in light and electron microscopy of strong objects, III: development of methods for numerical solution. *Opt. Acta*, 23:456–468, 1976.

[58] J. W. Wood, M. A. Fiddy, and R. E. Burge. Phase retrieval using two intensity measurements in the complex plane. *Opt. Let.*, 6(11):514–516, Nov 1981.

[59] G. Zheng, R. Horstmeyer, and C. Yang. Wide-field, high-resolution Fourier ptychographic microscopy. *Nat. Photonics*, 7(9):739–745, 2013.

Introduction to Nonnegative Matrix Factorization

Nicolas Gillis
Department of Mathematics and Operational Research
University of Mons, Rue de Houdain 9, 7000 Mons, Belgium
E-Mail: nicolas.gillis@umons.ac.be
https://sites.google.com/site/nicolasgillis

1 Introduction to constrained low-rank matrix approximations

Constrained low-rank matrix approximation (CLRMA) is becoming more and more popular because it is able to extract pertinent information from large data sets; see, for example, the recent survey [87]. CLRMA is equivalent to *linear dimensionality reduction*. Given a set of n data points $m_j \in \mathbb{R}^p$ $(j = 1, 2, \ldots, n)$, the goal is to find a set of r basis vectors $u_k \in \mathbb{R}^p$ $(k = 1, 2, \ldots, r)$ and the corresponding weights v_{kj} so that for all j, $m_j \approx \sum_{k=1}^{r} v_{kj}\, u_k$. This problem is equivalent to the low-rank approximation of matrix M, with

$$M = [m_1\ m_2\ \ldots\ m_n] \approx [u_1\ u_2\ \ldots\ u_k][v_1\ v_2\ \ldots\ v_n] = UV,$$

where each column of M is a data point, each column of U is a basis vector, and each column of V provides the coordinates of the corresponding column of M in the basis U. In other words, each column of M is approximated by a linear combination of the columns of U.

In practice, when dealing with such models, two key choices exist:

1. *Measure of the error $M-UV$.* Using the standard least-squares error, $\|M-UV\|_F^2 = \sum_{i,j}(M-UV)_{i,j}^2$, leads to principal component analysis (PCA) that can be solved by using the singular value decomposition (SVD). Surprisingly, one can show that the optimization problem in variables (U, V) has no spurious local minima (i.e., all local minima are global), which explains why it can be solved efficiently despite the error being nonconvex. Note that the resulting problem can be reformulated as a semidefinite program (SDP) by using the Ky Fan 2-k-Norm [29, Prop. 2.9].
   If data is missing or if weights are assigned to the entries of M, the problem can be cast as a weighted low-rank matrix approximation (WLRA) problem with error $\sum_{i,j} W_{i,j}(M-UV)_{i,j}^2$ for some nonnegative weight matrix W, where $W_{i,j} = 0$ when the entry (i, j) is missing [86]. Note that if W contains entries only in $\{0, 1\}$, then the problem is also referred to as PCA with missing data or low-rank matrix completion with noise.
   WLRA is widely used for recommender systems [61] that predict the preferences of users for a given product based on the product's attributes and user preferences.
   If the sum of the absolute values of the entries of the error $\sum_{i,j} |M-UV|_{i,j}$ is used, we obtain yet another variant more robust to outliers (sometimes referred to as robust PCA [15]). It can be used, for example, for background subtraction in video sequences where the noise (the moving objects) is assumed to be sparse while the background has low rank.
2. *Constraints that the factors U and V should satisfy.* These constraints depend on the application at hand and allow for meaningful interpretation of the factors. For example, k-means[1] is equivalent to requiring the factor V to have a single nonzero entry in each column that is equal to one, so that the columns of U are cluster centroids. Another widely used variant is sparse PCA, which requires that the factors (U and/or V) be sparse [28, 57, 69], thus yielding a more compact and easily interpretable decomposition (e.g., if V is sparse, each data point is the linear combination of only a few basis elements). Imposing componentwise nonnegativity on both factors U and V leads to nonnegative matrix factorization (NMF). For example, in document analysis where each column of M corresponds to a document (a vector of word counts), these nonnegativity constraints allow one to interpret the columns of the factor U as topics, and the columns of the factor V indicate in which proportion each document discusses each topic [64]. In this paper, we focus on this particular variant of CLRMA.

CLRMA problems are at the heart of many fields of applied mathematics and computer science, including, statistics and data analysis [56], machine learning and data mining [30], signal and image processing [1], graph theory [22], numerical linear algebra, and systems theory and control [72]. The good news for the optimization community is that these CLRMA models lead to a wide variety of theoretical and algorithmic challenges for optimizers: Can we solve these problems? Under which conditions? What is the most appropriate model for a given application? Which algorithm should we use in which situation? What type of guarantees can we provide?

CLRMA problems can be formulated in the following way:

$$\min_{U \in \Omega_U, V \in \Omega_V} \|M - UV\|. \tag{1}$$

As an introduction, below we discuss several aspects of (1).

Complexity. As soon as the norm $\|\cdot\|$ is not the Frobenius norm or the feasible domain has constraints (i.e., $\Omega_U \neq \mathbb{R}^{p\times r}$ or $\Omega_V \neq \mathbb{R}^{r\times n}$), the problem becomes difficult in most cases. For example, WLRA, robust PCA, NMF, and sparse PCA are all NP hard [44, 50, 90, 74]. An active direction of research is developing approximation algorithms for such problems; see, for example, [26] for the norm $\sum_{j=1}^{n} \|M(:,j) - UV(:,j)\|_2^p$ (for $p = 2$, this is PCA), [79] for WLRA, and [85] for the componentwise ℓ_1-norm.

Convexification. Under some conditions on the matrix M, convexification approaches can lead to optimality guarantees.

[1] k-means is the problem of finding a set of centroids u_k such that the sum of the distances between each data point and the closest centroid is minimized.

When there are no constraints ($\Omega_U = \mathbb{R}^{p\times r}$, $\Omega_V = \mathbb{R}^{r\times n}$), (1) can be equivalently rewritten as

$$\min_X \|M - X\| \text{ such that } \operatorname{rank}(X) = r.$$

From X, a solution (U, V) can be obtained by factorizing X (e.g., using the SVD). The most widely used convex models are based on minimizing the nuclear norm of X:

$$\min_X \|M - X\| + \lambda \|X\|_*, \tag{2}$$

where $\lambda > 0$ is a penalty parameter and $\|X\|_* = \sum_{i=1}^{\min(n,p)} \sigma_i(X) = \|\sigma(X)\|_1$, $\sigma(X)$ being the vector of singular values of X. This problem can be written as a semidefinite program; see [80] and the references therein.

When the matrix M satisfies some conditions depending on the model (in particular, M has to be close to a low-rank matrix), the optimal solution to (2) can be guaranteed to recover the solution of the original problem; examples include PCA with missing data [80] and robust PCA [19, 15].

As far as we know, these approaches have two drawbacks. First, if the input matrix M does not satisfy the required conditions, which is often the case in practice (e.g., for recommender systems and document classification where the input matrix is usually not close to a low-rank matrix), it is unclear whether the quality of the solution to (2) will be satisfactory. Second, the number of variables is much larger than in (1), namely, mn vs. $r(m+n)$. For large-scale problems, even first-order methods might be too costly. A possible way to handle the large positive semidefinite matrix is to (re)factor it in the SDP as the product of two matrices; this is sometimes referred to as the Burer-Monteiro approach [14]. In fact, in many cases, any stationary point can be guaranteed to be a global minimum [12, 66]; see also [65] for a survey. This is currently an active area of research: trying to identify nonconvex problems for which optimal solutions can be guaranteed to be computed efficiently (see the end of the next paragraph for other examples).

Nonconvex approaches. One can tackle (1) in many ways using standard nonlinear optimization schemes. The most straightforward and popular way is to use a two-block coordinate descent method (in particular if Ω_U and Ω_V are convex sets since the subproblems in U and V are convex):

0. Initialize (U, V).
1. $U \leftarrow X$, where X solves exactly or approximately $\min_{X\in\Omega_U} \|M - XV\|$.
2. $V \leftarrow Y$, where Y solves exactly or approximately $\operatorname{argmin}_{Y\in\Omega_V} \|M - UY\|$.

This simple scheme can be implemented in different ways. The subproblems are usually not solved up to high precision; for example, a few steps of a (fast) gradient method can be used. These methods can in general be guaranteed to converge to a stationary point of (1) [16]. More sophisticated schemes include Riemannian optimization techniques [11, 89]. Many methods based on randomization have also been developed recently; see the surveys [71, 91].

Alternating and local minimization were shown to lead to optimal solutions under assumptions similar to those needed for convexification-based approaches; see, for example, [59, 55] for PCA with missing data, [2] for (a variant of) sparse PCA, and [78] for robust PCA. Recently, [7, 38] showed that PCA with missing data has no spurious local minima (under appropriate conditions).

Outline of the paper. In the rest of this paper, we focus on a particular CLRMA problem, namely, nonnegative matrix factorization (NMF), with $\|\cdot\| = \|\cdot\|_F^2$, $\Omega_U = \mathbb{R}_+^{p\times r}$, and $\Omega_V = \mathbb{R}_+^{r\times n}$. As opposed to other CLRMA variants (such as robust PCA, sparse PCA, and PCA with missing data), as far as we know, no useful convexification approach exists.

The goal of this paper is not to provide an exhaustive survey but rather to provide a brief introduction, focusing only on several aspects of NMF (obviously biased toward our own interests). In particular, we address the application of NMF for hyperspectral imaging, the geometric interpretation of NMF, complexity issues, algorithms, and the nonnegative rank and its link with extended formulations of polyhedra.

2 Nonnegative Matrix Factorization

The standard NMF problem can be formulated as follows

$$\min_{U\in\mathbb{R}^{p\times r}, V\in\mathbb{R}^{r\times n}} \|M - UV\|_F^2 \text{ such that } U, V \geq 0. \tag{3}$$

As mentioned in the introduction, these nonnegativity constraints allow interpreting the basis elements in the same way as the data (e.g., as image, or vector of word counts) while the nonnegativity of V allows interpreting the weights as activation coefficients. We describe in detail in the next section a particular application, namely, blind hyperspectral unmixing, where the nonnegativity of U and V has a physical interpretation.

The nonnegativity constraints also naturally lead to sparse factors. In fact, the first-order optimality conditions of a problem of the type $\min_{x\geq 0} f(x)$ are $x_i \geq 0$, $\nabla_i f(x) \geq 0$ and $\nabla_i f(x) x_i = 0$ for all i. Hence stationary points of (3) are expected to have zero entries. This property of NMF enhances its interpretability and provides a better compression compared with unconstrained variants.

We refer to the problem of finding an exact factorization, that is, finding $U \geq 0$ and $V \geq 0$ such that $M = UV$, as "exact NMF." The minimum r such that an exact NMF exists is the nonnegative rank of M, denoted $\operatorname{rank}_+(M)$. We have that $\operatorname{rank}(M) \leq \operatorname{rank}_+(M) \leq \min(m, n)$ (since $M = MI = IM$, where I is the identify matrix).

NMF has been used successfully in many applications; see, for example, [25, 42] and the references therein. In the next section we focus on one particular application, namely, blind hyperspectral unmixing.

3 Hyperspectral Imaging

A grayscale image is an image in which the value of each pixel is a single sample. An RGB image has three channels (red, green, and blue) and allows a color image to be

reconstructed as it is perceived by an human eye. A hyperspectral image is an image for which usually each pixel has between 100 and 200 channels, corresponding to the reflectance (fraction of light reflected by that pixel) at different wavelengths. The wavelengths measured in a hyperspectral image depend on the camera used and are usually chosen depending on the application at hand. The advantage of hyperspectral images is that they contain much more information, some of it blind to the human eye, that allows one to identify and characterize the materials present in a scene much more precisely; see Figure 1 for an illustration. Its numerous applications include agriculture, eye care, food processing, mineralogy, surveillance, physics, astronomy, chemical imaging, and environmental science; see, for example, https://en.wikipedia.org/wiki/Hyperspectral_imaging or http://sciencenordic.com/lengthy-can-do-list-colour-camera.

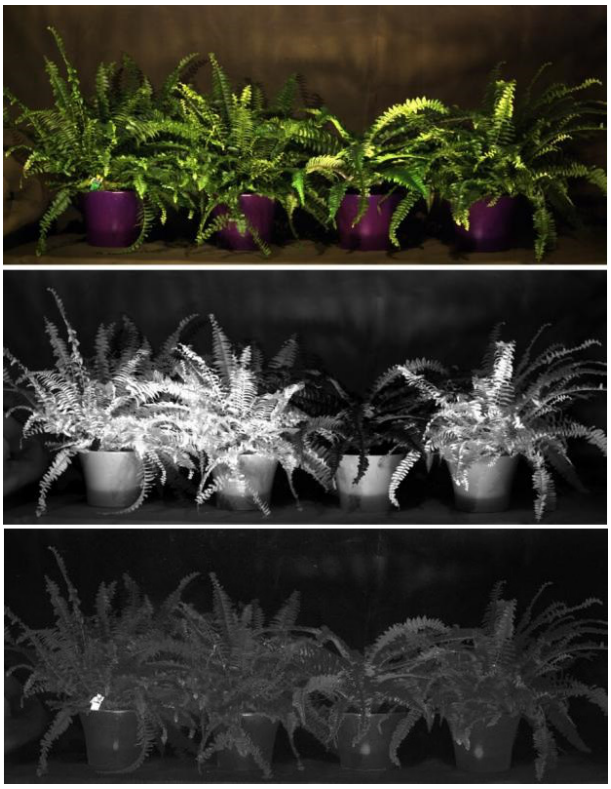


Figure 1: From top to bottom: (1) RGB image of four plants: can you identify the artificial one? (2) Grayscale image at a wavelength that is blind to the naked eye (namely, 770 nm, infrared) and allows identifying the artificial plant (plants have a high reflectance at infrared wavelengths, as opposed to the artificial material). (3) Analysis of the image allows finding a small target, a LEGO figure within the plants. Source: sciencenordic.com, Photo: Torbjørn Skauli, FFI.

Assume a scene is being imaged by a hyperspectral imager using p wavelengths (that is, p channels) and n pixels. Let us construct the matrix $M \in \mathbb{R}_+^{p\times n}$ such that $M(i,j)$ is the reflectance of the jth pixel at the ith wavelength. Each column of M therefore corresponds to the so-called spectral signature of a pixel, while each row corresponds to a vectorized image at a given wavelength. Given such an image, an important goal in practice is to (1) identify the constitutive materials present in the image, called endmembers (e.g., grass, trees, road surfaces, roof tops) and (2) classify the pixels accordingly, that is, identify which pixels contain which materials and in which quantity. In fact, the resolution of most hyperspectral images is low, and hence most pixels will contain several materials. If a library or dictionary of spectral signatures of materials present in the image is not available, this problem is referred to as blind hyperspectral unmixing (blind HU): the goal is to identify the endmembers and quantify the abundances of the endmembers in each pixel.

The simplest and most popular model is the linear mixing model (LMM). It assumes that the spectral signature of a pixel equals the weighted linear combination of the spectral signatures of the endmembers it contains, where the weight is given by the abundances. Physically, the reflectance of a pixel will be proportional to the materials it contains: for example, if a pixel contains 30% of aluminum and 70% of copper, its spectral signature will be equal to 0.3 times the spectral signature of the aluminum plus 0.7 times the spectral signature of the copper. In practice, this model is only approximate because of imperfect conditions (measurement noise, light reflecting off several times before being measured, atmospheric distortion, etc.). We refer the reader to [8, 70] for recent surveys on (blind) HU techniques and to [84] for an introduction to hyperspectral imaging.

If we use the LMM and assume that the image contains r endmembers whose spectral signatures are given by the columns of the matrix $U \in \mathbb{R}_+^{m\times r}$, we have for all j

$$M(:,j) = \sum_{k=1}^{r} v_{kj} U(:,k) = UV(:,j),$$

where $v_{kj} \geq 0$ is the abundance of the kth endmember in the jth pixel. Therefore, blind HU boils down to the NMF of matrix M; see Figure 2 for an illustration.

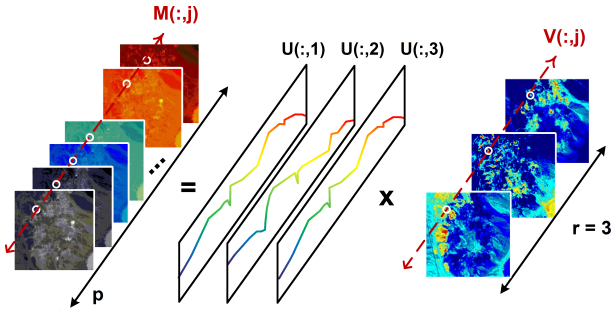


Figure 2: Illustration of the decomposition of a hyperspectral image with three endmembers [70]. On the left, the hyperspectral image M; in the middle, the spectral signatures of the three endmembers as the columns of matrix U; on the right, the abundances of each material in each pixel (referred to as the abundance maps).

Using a standard NMF algorithm, that is, an algorithm that tries to solve (3), will in general not lead to the sought decomposition. The reason is that the solution of

NMF is highly nonunique, as discussed later. In practice, a meaningful solution is achieved usually by using additional constraints/penalty terms, including: the sum-to-one constraints on the abundances ($\sum_{k=1}^{r} v_{kj} = 1 \, \forall j$), sparsity of V (because most pixels contain only a few endmembers), piecewise smoothness of the columns of U (since they correspond to spectral signatures), and spatial coherence of the rows of V (because neighboring pixels are more likely to contain the same endmembers). Numerous constrained variants of NMF exist that we do not discuss here; see, for example, [25, 42] and the references therein.

4 Geometry and Uniqueness

NMF has a nice geometric interpretation, which is crucial to consider in order to understand the nonuniqueness of the solutions. As discussed subsequently, it also allows one to develop efficient algorithms and is closely related to the extended formulations of polyhedra.

Let us consider the exact case, that is, $M = UV$. Without loss of generality, (i) the zero columns of M and U can be removed, and (ii) the columns of M and U can be normalized so that the entries of each column sum to one:

$$MD_M^{-1} = UD_U^{-1} D_U V D_M^{-1},$$

where D_M and D_U are diagonal matrices with $D_M(j,j) = \|M(:,j)\|_1$ and $D_U(j,j) = \|U(:,j)\|_1$, respectively. Since we have $M(:,j) = \sum_{k=1}^{r} U(:,k)V(k,j) = UV(:,j)$, this normalization implies that the columns of V also have their entries summing to one, that is, $\|V(:,j)\|_1 = 1$ for all j. Thus that, after normalization, the columns of M belong to the convex hull of the columns of U:

$$M(:,j) \in \text{conv}(U) \subseteq \Delta^p = \{x \in \mathbb{R}^p | x \geq 0, \|x\|_1 = 1\} \quad \forall j,$$

where $\text{conv}(U) = \{Ux | x \geq 0, \|x\|_1 = 1\}$. Therefore, the exact NMF problem is equivalent to finding a polytope, $\text{conv}(U)$, nested between two given polytopes, $\text{conv}(M)$ and the unit simplex Δ^p. The dimension of the inner polytope, $\text{conv}(M)$, is $\text{rank}(M) - 1$, while the dimension of the outer polytope, Δ^p, is $p-1$. The dimension of the nested polytope $\text{conv}(U)$ is not known in advance. When the three polytopes (inner, nested, and outer) have the same dimension, this problem is well known in computational geometry and is referred to as the nested polytope problem (NPP) [27].

If $\text{rank}(M) = \text{rank}(U)$, the column spaces of M and U must coincide, and the outer polytope can be restricted to $\Delta^p \cap \text{col}(M)$, in which case the inner, nested, and outer polytopes have the same dimension. If we impose explicitly this additional constraint ($\text{rank}(M) = \text{rank}(U)$) on the exact NMF problem, we can prove that NPP and this restricted variant of exact NMF are equivalent, that is, they can be reduced to one another [46, 20].

To illustrate, we present a simple example with nested hexagons; this is similar to the example presented in [76]. Let $a > 1$, and let M_a be the matrix

$$\frac{1}{a} \begin{pmatrix} 1 & a & 2a-1 & 2a-1 & a & 1 \\ 1 & 1 & a & 2a-1 & 2a-1 & a \\ a & 1 & 1 & a & 2a-1 & 2a-1 \\ 2a-1 & a & 1 & 1 & a & 2a-1 \\ 2a-1 & 2a-1 & a & 1 & 1 & a \\ a & 2a-1 & 2a-1 & a & 1 & 1 \end{pmatrix}. \tag{4}$$

The restricted exact NMF problem for M_a involves two nested hexagons (recall that we restrict the polytopes to be in the intersection between the column space of M_a and Δ^p, which has dimension 2 since $\text{rank}(M_a) = 3$). Each facet of the outer polytope corresponds to a facet of the nonnegative orthant, that is, to a nonnegativity constraint. The inner hexagon is smaller than the outer one with a ratio of $\frac{a-1}{a}$.

For $a = 2$, the inner hexagon is twice as small as the outer one, and we can fit a triangle between the two so that $\text{rank}_+(M_a) = 3$; see Figure 3 (top). For any $a > 2$, $\text{rank}_+(M_a) \geq 4$ because no triangle can fit between the two hexagons. For $a = 3$, the inner hexagon is 2/3 smaller than the outer one, and we can fit a rectangle between the two and $\text{rank}_+(M_a) = 4$; see Figure 3 (bottom). This implies that $\text{rank}_+(M_a) = 4$ for all $2 < a \leq 3$.

For any $a > 3$, $\text{rank}_+(M_a) = 5$. Surprisingly, the nonnegative rank of M_a is always no more than 5 (even when a tends to infinity, in which case the inner and outer hexagons coincide) because there exists a three-dimensional polytope within Δ^6 with 5 vertices that contains the outer polytope; see Figure 4, which corresponds to the factorization

$$\begin{aligned} M = \lim_{a \to +\infty} M_a &= \begin{pmatrix} 0 & 1 & 2 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 2 & 1 \\ 1 & 0 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 0 & 1 & 2 \\ 2 & 2 & 1 & 0 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 & 0 \end{pmatrix} \\ &= UV \\ &= \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 2 \\ 0 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}, \end{aligned} \tag{5}$$

where $\text{rank}(U) = 4$, and hence $\text{conv}(U)$ has dimension 3.

This example illustrates other interesting properties of NMF:

- NMF does not in general have a unique solution (up to scaling and permutation of the rank-one factors). For example, for $a = 2$ (Figure 3, top), four triangles can be fit between the two polytopes (the one shown on the figure, its rotation by 60 degrees, and two triangles whose vertices are three nonadjacent vertices of the outer hexagon). For $1 < a < 2$, this would be even worse since there would be an infinite number of solutions. For this reason, practitioners often add additional constraints to the NMF model to try to identify the most meaningful solution to their problem (such as sparsity,

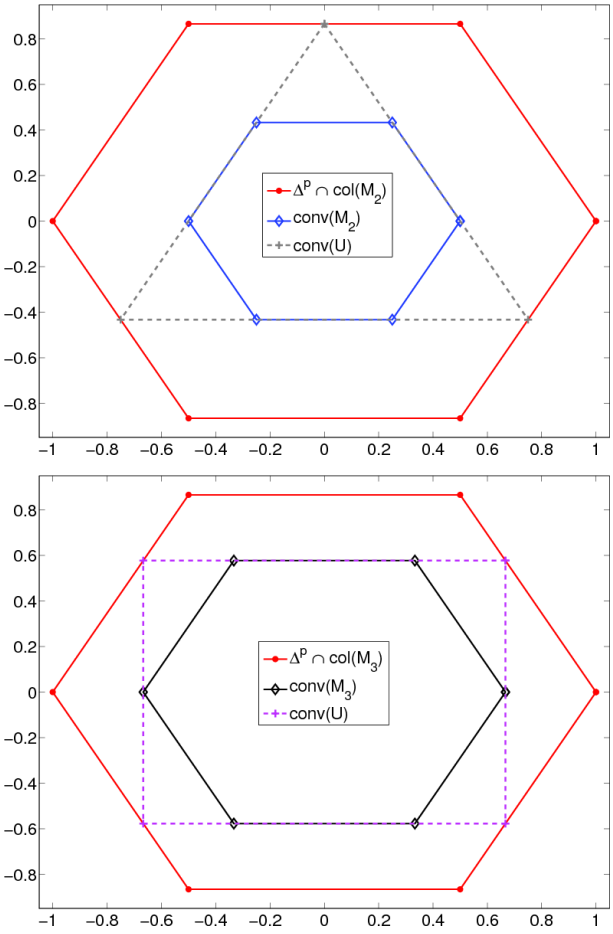


Figure 3: NPP problem corresponding to the exact NMF of the matrix from (4), restricted to the column space of M: (top) the case $a = 2$; (bottom) $a = 3$.

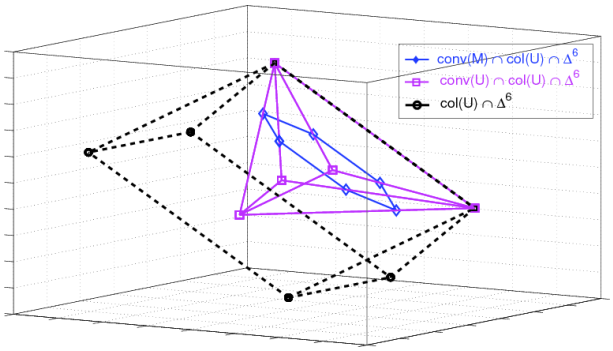


Figure 4: NPP solution corresponding to the exact NMF of the matrix from (5), restricted to the column space of U. It corresponds to the matrix M_a from (4) when $a \to \infty$.

as discussed earlier); see, for example, [63, 40, 54] for more details on the uniqueness of NMF.

- The nonnegative rank can increase only in the neighborhood of a given matrix; that is, the nonnegative rank is upper semicontinuous [10, Th.3.1]:

  If P is a nonnegative matrix, without zero columns and with $\text{rank}_+(P) = r$, then there exists a ball $B(P, \epsilon)$ centered at P and of radius $\epsilon > 0$ such that $\text{rank}_+(N) \geq r$ for all $N \in B(P, e)$.

5 Complexity

Given a nonnegative matrix M, checking whether $\text{rank}(M) = \text{rank}_+(M) = r$ is NP hard: unless $P = NP$, there is no polynomial time algorithm in m, n and r for this problem [90]. If r is fixed, however, there is a polynomial time algorithm running in $O\big((pn)^{r^2}\big)$ [5, 75]. The argument is based on quantifier elimination theory (in particular the fact that checking whether a system of ℓ equations in n variables up to degree d can be solved in time polynomial in ℓ and d). Unfortunately, as far as we know, this cannot be used in practice, even for small matrices (e.g., checking whether a 4-by-4 matrix has nonnegative rank 3 seems already impractical with current solvers). Developing an effective code for exact NMF for small matrices is an important direction for further research. Note that we have developed a code based on heuristics that allows solving exact NMF for matrices up to a few dozen rows and columns (although our code comes with no guarantee) [88].

More recently, Shitov [83] and independently Chistikov et al. [21] answered an important open problem showing that the nonnegative rank over the reals might be different from the nonnegative rank over the rationals, implying that the nonnegative rank computation is not in NP since the size of the output is not bounded by the size of the input.

6 Algorithms

In this section, we briefly describe the two main classes of NMF algorithms. As mentioned in the introduction, there does not exist, to the best of our knowledge, a successful convexification approach for NMF, as opposed to other low-rank models. Note, however, that there does exist a convexification approach to compute lower bounds for the nonnegative rank [35]. An explanation is that we cannot work directly with the low-rank approximation $X = UV$ and use the nuclear norm of X, because even if we were given the best nonnegative approximation X of nonnegative rank r for M, in general recovering the exact NMF (U, V) of X would be difficult. Writing directly a convexification in variables (U, V) seems difficult (for rank higher than one[2]) because of the symmetry of the problem (permuting columns of U and rows of V accordingly provides an equivalent solution). Breaking this symmetry seems nontrivial; see [39, pp. 146-148] for a discussion and a tentative SDP formulation. This is an interesting direction for further research.

6.1 Standard nonlinear optimization schemes

As for CLRMA problems, most NMF algorithms use a two-block coordinate descent scheme:

0. Initialize $(U, V) \geq 0$.

[2]Note that the rank-one NMF problem is equivalent to the rank-one unconstrained problem since for any rank-one solution uv^T, one can easily check that $|u||v|^T$ is a solution with lower objective function value. This also follows from the Perron-Frobenius and Eckart-Young theorems.

1. $U \leftarrow X$, where X solves exactly or approximately $\min_{X \geq 0} \|M - XV\|_F$.
2. $V \leftarrow Y$, where Y solves exactly or approximately $\operatorname{argmin}_{Y \geq 0} \|M - UY\|_F$.

Note that the subproblems to be solved are so-called nonnegative least squares (NNLS). Because NMF is NP hard, these algorithms can only guarantee convergence (usually to a first-order stationary point).

The most well-known algorithm for NMF is the multiplicative updates, namely,

$$U \leftarrow U \circ \frac{[MV^T]}{[UVV^T]}, \quad V \leftarrow V \circ \frac{[U^TM]}{[U^TUV]},$$

where $\circ$ (resp. $\frac{[]}{[]}$) is the componentwise product (resp. division) between two matrices. It is extremely popular because of its simplicity and because it was proposed in the paper of Lee and Seung [64] that launched the research on NMF. However, it converges slowly; it cannot modify zero entries; and it is not guaranteed to converge to a stationary point. Note that it can be interpreted as a rescaled gradient descent; see, for example, [42].

Methods that try to solve the subproblems exactly are referred to as alternating nonnegative least squares; among these, active set methods seem to be the most efficient, and dedicated codes have been implemented by Haesun Park and collaborators; see [60] and the references therein.

In practice, a method that seems to work extremely well is to apply a few steps of coordinate descent on the NNLS subproblems: the subblocks are the columns of U and the rows of V [24, 45]—the reason is that the subproblems can be solved in closed form. In fact, the optimal kth column of U (all other variables being fixed) is given by

$$\arg \min_{U(:,k) \geq 0} \|R_k - U(:,k)V(k,:)\|_F^2 = \max\left(0, \frac{R_kV(k,:)^T}{\|V(k,:)\|_2^2}\right),$$

for $R_k = M - \sum_{j \neq k} U(:,j)V(j,:)$, and similarly by symmetry for the kth row of V.

Many other approaches can be applied to the NNLS subproblems (e.g., projected gradient method [67], fast/accelerated gradient method (Nesterov's method) [53], and Newton-like method [23]).

6.2 Separable NMF

Although they usually provide satisfactory results in practice, the methods described in the preceding section do not come with any guarantee. In their paper on the complexity of NMF, Arora et al. [5] also identify a subclass of matrices for which the NMF problem is much easier. These are the so-called separable matrices defined as follows.

Definition 1. *A matrix M is separable if there exists a subset $\mathcal{K}$ of r of its columns with $r = \operatorname{rank}_+(M)$ and a nonnegative matrix V such that $M = M(:,\mathcal{K})V$.*

This requires each column of the basis matrix U in an NMF decomposition to be present in the input matrix M. Equivalently, this requires the matrix V in an NMF decomposition to contain the identity matrix as a submatrix. The separable NMF problem is the problem to identify the subset $\mathcal{K}$ (in the noisy case, this subset should be such that $\min_{V \geq 0} \|M - M(:,\mathcal{K})V\|$ is minimized).

Although this condition is strong, it makes sense in several applications, for example the following.

- Document classification: for each topic, there is a "pure" word used only by that topic (an "anchor" word) [4].
- Time-resolved Raman spectra analysis: each substance has a peak in its spectrum while the other spectra are (close to) zero [68].
- Blind hyperspectral unmixing: for each endmember, there exists a pixel that contains only that endmember. This is the so-called pure-pixel assumption that has been used since the 1990s in that community.

Other applications include video summarization [31] and foreground-background separation [62].

Geometrically, in the exact case and after normalization of the columns of X and U, the separability assumption is equivalent to having $\operatorname{conv}(U) = \operatorname{conv}(M)$. Therefore, the so-called separable NMF problem reduces to identify the vertices of the convex hull of the columns of M. This is a relatively easy geometric problem. It becomes tricky when noise is added to the separable matrix, and many recent works have tried to quantify the level of noise that one can tolerate and still be able to recover the vertices, up to some error.

Geometric algorithms

Most algorithms for separable NMF are based on the geometric interpretation, many being developed within the blind HU community (sometimes referred to as pure-pixel search algorithms). Only recently, however, has robustness to noise of these algorithms been analyzed.

One of the simplest algorithm, often referred to as the successive projection algorithm, is closely related to the modified Gram-Schmidt algorithm with column pivoting and has been discovered several times [3, 81, 18]; see the discussion in [70]. Over a polytope, a strongly convex function (such as the ℓ_2 norm) is always maximized at a vertex: this can be used to identify a vertex, that is, a column of U (recall that we assume that the columns of M are normalized so that $\operatorname{conv}(U) = \operatorname{conv}(M)$ under the separability assumption). Once a column of U has been identified, one can project all columns of M onto the orthogonal complement of that column (so that this particular column projects onto 0): this amounts to applying a linear transformation to the polytope. If U is full rank (meaning the polytope is a simplex, which is the case usually in practice), then the other vertices do not project onto 0, and one can use these two steps recursively. This approach is a greedy method to identify a subset of the columns with maximum volume [17, 18]. This algorithm was proved to be robust to noise [49] and can be made more robust to noise by using strategies such as

- applying dimensionality reduction, such as PCA, to the columns of M in order to filter the noise [77];
- using a preconditioning based on minimum-volume el-

lipsoid [43, 73];
- going over the identified vertices (once r vertices have been identified) to check whether they still maximize the strongly convex function once projected onto the orthogonal complement of the other vertices (otherwise, they are replaced, increasing the volume of the identified vertices) [4]; and
- taking into account the nonnegativity constraints in the projection step [41].

We refer the reader to [8, 70] for surveys on these approaches. Most geometric approaches for separable NMF are computationally cheap. Usually, however, they are sensitive to outliers.

Convex models

If M is separable, there exist an index set $\mathcal{K}$ of size r and a nonnegative matrix V such that $M = M(:, \mathcal{K})V$. Equivalently, there exists an n-by-n nonnegative matrix X with r nonzero rows such that $M = MX$ with $X(\mathcal{K}, :) = V$. Solving separable NMF can therefore be formulated as

$$\min_{X \geq 0} \|X\|_{\text{row},0} \quad \text{such that } M = MX,$$

where $\|X\|_{\text{row},0}$ counts the number of nonzero rows of X. A standard convexification approach is to use the ℓ_1 norm, replacing $\|X\|_{\text{row},0}$ with $\sum_{i=1}^{n} \|X(i,:)\|_k$ for some k; for example, [31] uses $k = \infty$ and [32] uses $k = 2$.

If the columns of M are normalized, the entries of V are bounded above by one (since the columns of U are vertices), and another formulation for separable NMF is obtained:

$$\begin{array}{ll} \min_{X \geq 0} & \|\operatorname{diag}(X)\|_0 \\ \text{such that} & M = MX \text{ and } X(i,j) \leq X(i,i) \leq 1 \quad \forall i, j. \end{array}$$

Because on each row the diagonal entry has to be the largest and because the goal is to minimize the number of nonzero entries of the diagonal of X, the optimal solution will contain r nonzero diagonal entries and hence r nonzero rows. (Note that requiring the diagonal entries of X to be binary would allow one to model this problem exactly by using mixed-integer linear programming.) Using the ℓ_1 norm, we get another convex model (proposed in [9] and improved in [47]):

$$\begin{array}{ll} \min_{X \geq 0} & \operatorname{trace}(X) \\ \text{such that} & M = MX \text{ and } X(i,j) \leq X(i,i) \leq 1 \quad \forall i, j, \end{array}$$

where $\operatorname{trace}(X)$ is equal to $\|\operatorname{diag}(X)\|_1$ since X is nonnegative. In practice, when noise is present, the equality term $M = MX$ is replaced with $\|M - MX\| \leq \epsilon$ for some appropriate norm (typically the ℓ_1, ℓ_2, or Frobenius norm) or is added in the objective function as a penalty.

The two models presented above turn out to be essentially equivalent [48]. The main drawback is the computational cost, since these models have n^2 variables. For example, in hyperspectral imaging, n is the number of pixels and is typically on the order of millions; hence, solving these problems is challenging (if not impractical). A natural approach is therefore to first select a subset of good candidates among the columns of M (e.g., using geometric algorithms) and then optimize only over this subset of the rows of X [32, 48]. The main advantage of this approach is that the resulting models are provably the most robust for separable NMF [47]. Intuitively, the reason is not only that the model focuses in identifying, for example, a subset of columns with large volume but also that it requires all the data points to be well approximated with the selected vertices (since $\|M - MX\|$ should be small). For this reason, they are also much less sensitive to outliers than are most geometric approaches.

7 Nonnegative Rank and Extended Formulations

We now describe the link between extended formulations of polyhedra and NMF. This is closely related to the geometric interpretation of NMF described earlier.

Let $\mathcal{P}$ be a polytope

$$\mathcal{P} = \{x \in \mathbb{R}^k \mid b_i - A(i,:)x \geq 0 \text{ for } 1 \leq i \leq p\},$$

and let $(w_1, \cdots, w_n)$ be its vertices. Let $S_{\mathcal{P}}$ be the p-by-n slack matrix of $\mathcal{P}$ defined as follows:

$$S_{\mathcal{P}}(i,j) = b_i - A(i,:)w_j \qquad 1 \leq i \leq p,\ 1 \leq j \leq n.$$

An extended formulation of $\mathcal{P}$ is a higher-dimensional polyhedron $Q \subseteq \mathbb{R}^{k+p}$ that (linearly) projects onto P. The minimum number of facets (that is, inequalities) of such a polytope is called the extension complexity, $\operatorname{xp}(\mathcal{P})$, of $\mathcal{P}$.

Theorem 1. *(Yannakakis, [92]). Let $S_{\mathcal{P}}$ be the slack matrix of the polytope $\mathcal{P}$. Then,* $\operatorname{rank}_+(S_{\mathcal{P}}) = \operatorname{xp}(\mathcal{P})$.

Let us just show that $\operatorname{xp}(\mathcal{P}) \leq \operatorname{rank}_+(S_{\mathcal{P}})$, because it is elegant and straightforward. Given $\mathcal{P} = \{x \in \mathbb{R}^k \mid b - Ax \geq 0\}$, any exact NMF of $S_{\mathcal{P}} = UV$ with $U \geq 0$ and $V \geq 0$ provides an explicit extended formulation (with some redundant equalities) of $\mathcal{P}$:

$$Q = \{(x,y) \mid b - Ax = Uy \text{ and } y \geq 0\}.$$

In fact, let us show that $Q_x = \{x | \exists y \text{ s.t. } (x,y) \in Q\} = \mathcal{P}$. We have $Q_x \subseteq \mathcal{P}$ since $U \geq 0$ and $y \geq 0$; hence $b - Ax = Uy \geq 0$ for all $(x,y) \in Q$. We have $\mathcal{P} \subseteq Q_x$ because all vertices of $\mathcal{P}$ belong to Q_x: by construction, $(w_j, V(:,j)) \in Q$ since $S_{\mathcal{P}}(:,j) = b - Aw_j = UV(:,j)$ and $V(:,j) \geq 0$.

Example. The extension complexity of the regular n-polygons is $O(\log_2(n))$ [37]. This result can be used to approximate a second-order cone program with a linear program [6]. In particular, we have seen that the extension complexity of the regular hexagon is 5; see Equation (5) and Figure 4.

Recent results. Several recent important results for understanding the limits of linear programming for solving combinatorial problems are based on Theorem 1 and on constructing lower bounds for the nonnegative rank, usually based on the sparsity pattern of the slack matrix [36]; see [58] for a

survey. In particular, Rothvoß showed recently that the prefect matching problem cannot be written with polynomially many constraints [82].

These ideas can be generalized in two ways:

- To characterize the size of approximate extended formulations (for a given precision) [13].
- To any convex cone [51], which leads to other CLRMA problems. For example, for the cone of positive semidefinite (PSD) matrices, the rows of U and the columns of V are required to be vectorized PSD matrices. The smallest PSD extension of a given set (e.g., a polyhedron) is equal to the so-called PSD rank of its slack matrix; see the recent survey [34]. (Note that for non-polyhedral sets, the slack matrix is infinite since the number of extreme points and facets is not finite.)

These ideas, for example, recently allowed Hamza Fawzi to prove that the PSD cone cannot be represented using the second-order cone [33]; the proof relies on the fact that the second-order cone rank of the cone of 3-by-3 PSD matrices is infinite.

8 Conclusion

In this paper, we have introduced the NMF problem and discussed several of its aspects. The opportunity for meaningful interpretations is the main reason why NMF became so popular and has been used in many applications. NMF is tightly connected with difficult geometric problems; hence developing fast and reliable algorithms is a challenge. Although important challenges remain to be tackled (e.g., developing exact algorithms for small-scale problems), even more challenges exist in generalizations of NMF. In particular, we mentioned cone factorizations (such as the PSD factorization and its symmetric variant [52]), which are more recent problems and have not been explored to their full extent.

Acknowledgments. NG acknowledges the support by the F.R.S.-FNRS (incentive grant for scientific research n° F.4501.16) and by the ERC (starting grant n° 679515).

REFERENCES

[1] Special issue on source separation and application, IEEE Signal Process. Mag., 2014. http://online.qmags.com/SIPR0514.

[2] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used overcomplete dictionaries. In *COLT*, pages 123–137, 2014.

[3] U. Araújo, B. Saldanha, R. Galvão, T. Yoneyama, H. Chame, and V. Visani. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometr. Intell. Lab.*, 57(2):65–73, 2001.

[4] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *Int. Conf. on Machine Learning (ICML '13)*, volume 28, pages 280–288. 2013.

[5] S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization – provably. In *Proc. of the 44th Symp. on Theory of Computing (STOC '12)*, pages 145–162, 2012.

[6] A. Ben-Tal and A. Nemirovski. On polyhedral approximations of the second-order cone. *Math. Oper. Res.*, 26(2):193–205, 2001.

[7] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. *arXiv:1605.07221*, 2016.

[8] J. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.*, 5(2):354–379, 2012.

[9] V. Bittorf, B. Recht, E. Ré, and J. Tropp. Factoring nonnegative matrices with linear programs. In *Adv. Neur. In. (NIPS)*, pages 1223–1231, 2012.

[10] C. Bocci, E. Carlini, and F. Rapallo. Perturbation of matrices and nonnegative rank with a view toward statistical models. *SIAM J. Matrix Anal. Appl.*, 32(4):1500–1512, 2011.

[11] N. Boumal and P.-A. Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *Adv. Neur. In. (NIPS)*, pages 406–414, 2011.

[12] N. Boumal, V. Voroninski, and A. Bandeira. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. *arXiv:1606.04970*, 2016.

[13] G. Braun, S. Fiorini, S. Pokutta, and D. Steurer. Approximation limits of linear programs (beyond hierarchies). *Math. Oper. Res.*, 40(3):756–772, 2015.

[14] S. Burer and R. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Math. Program.*, 95(2):329–357, 2003.

[15] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11, 2011.

[16] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[17] A. Çivril and M. Magdon-Ismail. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theor. Comput. Sci.*, 410(47–49):4801–4811, 2009.

[18] T.-H. Chan, W.-K. Ma, A. Ambikapathi, and C.-Y. Chi. A simplex volume maximization framework for hyperspectral endmember extraction. *IEEE Trans. Geosci. Remote Sens.*, 49(11):4177–4193, 2011.

[19] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.*, 21(2):572–596, 2011.

[20] D. Chistikov, S. Kiefer, I. Marušić, M. Shirmohammadi, and J. Worrell. On restricted nonnegative matrix factorization. In *Proceedings of the 43rd International Colloquium on Automata Languages and Programming (ICALP)*, 2016.

[21] D. Chistikov, S. Kiefer, I. Marušić, M. Shirmohammadi, and J. Worrell. On rationality of nonnegative matrix factorization. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2017.

[22] F. Chung. *Spectral Graph Theory*, volume 92. American Mathematical Soc., 1997.

[23] A. Cichocki, R. Zdunek, and S.-I. Amari. Non-negative matrix factorization with quasi-Newton optimization. In *Lecture Notes in Artificial Intelligence, Springer*, volume 4029, pages 870–879, 2006.

[24] A. Cichocki, R. Zdunek, and S.-I. Amari. Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. In *Lecture Notes in Computer Science, Springer, vol. 4666*, pages 169–176, 2007.

[25] A. Cichocki, R. Zdunek, A. Phan, and S.-I. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation.* John Wiley & Sons, 2009.

[26] K. Clarkson and D. Woodruff. Input sparsity and hardness for robust subspace approximation. In *56th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2015)*, 2015.

[27] G. Das and D. Joseph. The complexity of minimum convex nested polyhedra. In *Proceedings of the 2nd Canadian Conference on Computational Geometry*, pages 296–301, 1990.

[28] A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.*, 49(3):434–448, 2007.

[29] X. Doan and S. Vavasis. Finding the largest low-rank clusters with Ky Fan 2-k-norm and ℓ_1-norm. *SIAM J. Optim.*, 26(1):274–312, 2016.

[30] L. Eldén. *Matrix Methods in Data Mining and Pattern Recognition*, volume 4. SIAM, 2007.

[31] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '12)*, 2012.

[32] E. Esser, M. Moller, S. Osher, G. Sapiro, and J. Xin. A convex model for nonnegative matrix factorization and dimensionality reduction on physical space. *IEEE Trans. Image Process.*, 21(7):3239–3252, 2012.

[33] H. Fawzi. On representing the positive semidefinite cone using the second-order cone. *arXiv:1610.04901*, 2016.

[34] H. Fawzi, J. Gouveia, P. Parrilo, R. Robinson, and R. Thomas. Positive semidefinite rank. *Math. Program.*, 153(1):133–177, 2015.

[35] H. Fawzi and P. Parrilo. Lower bounds on nonnegative rank via nonnegative nuclear norms. *Math. Program.*, 153(1):41–66, 2015.

[36] S. Fiorini, V. Kaibel, K. Pashkovich, and D. Theis. Combinatorial bounds on nonnegative rank and extended formulations. *Discrete Math.*, 313(1):67–83, 2013.

[37] S. Fiorini, T. Rothvoss, and H. Tiwary. Extended formulations for polygons. *Discrete Comput. Geom.*, 48(3):658–668, 2012.

[38] R. Ge, J. Lee, and T. Ma. Matrix completion has no spurious local minimum. *arXiv:1605.07272*, 2016.

[39] N. Gillis. *Nonnegative Matrix Factorization: Complexity, Algorithms and Applications*. PhD thesis, Université Catholique de Louvain, 2011. https://sites.google.com/site/nicolasgillis/.

[40] N. Gillis. Sparse and unique nonnegative matrix factorization through data preprocessing. *Journal of Machine Learning Research*, 13(Nov):3349–3386, 2012.

[41] N. Gillis. Successive nonnegative projection algorithm for robust nonnegative blind source separation. *SIAM J. Imaging Sci.*, 7(2):1420–1450, 2014.

[42] N. Gillis. The why and how of nonnegative matrix factorization. In J. Suykens, M. Signoretto, and A. Argyriou, editors, *Regularization, Optimization, Kernels, and Support Vector Machines*, pages 257–291. Chapman & Hall/CRC, Machine Learning and Pattern Recognition Series, 2014.

[43] N. Gillis and S. A. Vavasis. Semidefinite programming based preconditioning for more robust near-separable nonnegative matrix factorization. *SIAM J. Optim.*, 25(1):677–698, 2015.

[44] N. Gillis and F. Glineur. Low-rank matrix approximation with weights or missing data is NP-hard. *SIAM J. Matrix Anal. Appl.*, 32(4):1149–1165, 2011.

[45] N. Gillis and F. Glineur. Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. *Neural Comput.*, 24(4):1085–1105, 2012.

[46] N. Gillis and F. Glineur. On the geometric interpretation of the nonnegative rank. *Linear Algebra Appl.*, 437(11):2685–2712, 2012.

[47] N. Gillis and R. Luce. Robust near-separable nonnegative matrix factorization using linear optimization. *J. Mach. Learn. Res.*, 15(1):1249–1280, 2014.

[48] N. Gillis and R. Luce. A fast gradient method for nonnegative sparse regression with self dictionary. *arXiv:1610.01349*, 2016.

[49] N. Gillis and S. Vavasis. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(4), 2014.

[50] N. Gillis and S. Vavasis. On the complexity of robust PCA and ℓ_1-norm low-rank matrix approximation. *arXiv:1509.09236*, 2015.

[51] J. Gouveia, P. A. Parrilo, and R. R. Thomas. Lifts of convex sets and cone factorizations. *Math. Oper. Res.*, 38(2):248–264, 2013.

[52] S. Gribling, D. de Laat, and M. Laurent. Matrices with high completely positive semidefinite rank. *Linear Algebra Appl.*, 513:122–148, 2017.

[53] N. Guan, D. Tao, Z. Luo, and B. Yuan. NeNMF: an optimal gradient method for nonnegative matrix factorization. *IEEE Trans. Signal Process.*, 60(6):2882–2898, 2012.

[54] K. Huang, N. Sidiropoulos, and A. Swami. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Trans. Signal Process.*, 62(1):211–224, 2014.

[55] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the 45th annual ACM Symposium on Theory of Computing*, pages 665–674. ACM, 2013.

[56] I. Jolliffe. *Principal Component Analysis*. Springer, second edition, 2002.

[57] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.*, 11(Feb.):517–553, 2010.

[58] V. Kaibel. Extended formulations in combinatorial optimization. *Optima*, 85:2–7, 2011.

[59] R. Keshavan, S. Oh, and A. Montanari. Matrix completion from a few entries. In *2009 IEEE International Symposium on Information Theory*, pages 324–328. IEEE, 2009.

[60] J. Kim, Y. He, and H. Park. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, 2014.

[61] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

[62] A. Kumar and V. Sindhwani. Near-separable non-negative matrix factorization with l1 and Bregman loss functions. In *Proceedings of the 2015 SIAM Int. Conf. on Data Mining*, pages 343–351, 2015.

[63] H. Laurberg, M. Christensen, M. Plumbley, L. Hansen, and S. Jensen. Theorems on positive data: On the uniqueness of NMF. *Computational Intelligence and Neuroscience*, 2008, 2008.

[64] D. Lee and H. Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[65] A. Lemon, A. M.-C. So, and Y. Ye. Low-rank semidefinite programming: Theory and applications. *Foundations and Trends in Optimization*, 2(1-2):1–156, 2016.

[66] Q. Li and G. Tang. The nonconvex geometry of low-rank matrix optimizations with general objective functions. *arXiv:1611.03060*, 2016.

[67] C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Comput.*, 19(10):2756–2779, 2007.

[68] R. Luce, P. Hildebrandt, U. Kuhlmann, and J. Liesen. Using separable non-negative matrix factorization techniques for the analysis of time-resolved Raman spectra. *Appl. Spectrosc.*, 70(9):1464–1475, 2016.

[69] R. Luss and M. Teboulle. Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint. *SIAM Rev.*, 55(1):65–98, 2013.

[70] W.-K. Ma, J. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. Plaza, A. Ambikapathi, and C.-Y. Chi. A signal processing perspective on hyperspectral unmixing. *IEEE Signal Process. Mag.*, 31(1):67–81, 2014.

[71] M. W. Mahoney. Randomized algorithms for matrices and data. *Found. and Trends in Mach. Learn.*, 3(2):123–224, 2011.

[72] I. Markovsky. *Low Rank Approximation: Algorithms, Implementation, Applications*. Springer Science & Business Media, 2011.

[73] T. Mizutani. Robustness analysis of preconditioned successive projection algorithm for general form of separable NMF problem. *Linear Algebra Appl.*, 497:1–22, 2016.

[74] B. Moghaddam, Y. Weiss, and S. Avidan. Generalized spectral bounds for sparse LDA. In *Proceedings of the 23rd international conference on Machine learning*, pages 641–648. ACM, 2006.

[75] A. Moitra. An almost optimal algorithm for computing nonnegative rank. In *Proc. of the 24th Annual ACM-SIAM Symp. on Discrete Algorithms (SODA '13)*, pages 1454–1464, 2013.

[76] D. Mond, J. Smith, and D. Van Straten. Stochastic factorizations, sandwiched simplices and the topology of the space of explanations. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 459, pages 2821–2845. The Royal Society, 2003.

[77] J. Nascimento and J. Bioucas-Dias. Vertex component analysis: a fast algorithm to unmix hyperspectral data. *IEEE Trans. Geosci. Remote Sens.*, 43(4):898–910, 2005.

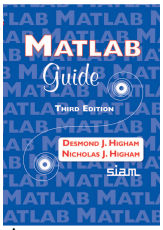
[78] P. Netrapalli, U. Niranjan, S. Sanghavi, and A. Anandkumar. Non-convex robust PCA. In *Adv. Neur. In. (NIPS)*, pages 2080–2088, 2014.

[79] I. Razenshteyn, Z. Song, and D. P. Woodruff. Weighted low rank approximations with provable guarantees. In *Proc. of the 48th Symp. on Theory of Computing (STOC '16)*, pages 250–263, 2016.
[80] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.
[81] H. Ren and C.-I. Chang. Automatic spectral target recognition in hyperspectral imagery. *IEEE Trans. Aerosp. Electron. Syst.*, 39(4):1232–1249, 2003.
[82] T. Rothvoß. The matching polytope has exponential extension complexity. In *Proceedings of the 46th annual ACM Symposium on Theory of Computing*, pages 263–272. ACM, 2014.
[83] Y. Shitov. Nonnegative rank depends on the field II. *arXiv:1605.07173*, 2016.
[84] R. Smith. Introduction to hyperspectral imaging. 2006. Microimages, http://www.microimages.com/documentation/Tutorials/hyprspec.pdf.
[85] Z. Song, D. P. Woodruff, and P. Zhong. Low rank approximation with entrywise ℓ_1-norm error. *arXiv:1611.00898*, 2016.
[86] N. Srebro and T. Jaakkola. Weighted low-rank approximations. In *ICML*, volume 3, pages 720–727, 2003.
[87] M. Udell, C. Horn, R. Zadeh, and S. Boyd. Generalized low rank models. *Found. and Trends in Mach. Learn.*, 9(1):1–118, 2016.
[88] A. Vandaele, N. Gillis, F. Glineur, and D. Tuyttens. Heuristics for exact nonnegative matrix factorization. *J. of Global Optim.*, 65(2):369–400, 2016.
[89] B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM J. Optim.*, 23(2):1214–1236, 2013.
[90] S. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM J. Optim.*, 20(3):1364–1377, 2010.
[91] D. Woodruf. Sketching as a tool for numerical linear algebra. *Found. and Trends in Mach. Learn.*, 10(1-2):1–157, 2014.
[92] M. Yannakakis. Expressing combinatorial optimization problems by linear programs. *J. Comput. Syst. Sci.*, 43(3):441–466, 1991.

Bulletin

1 Book Announcements

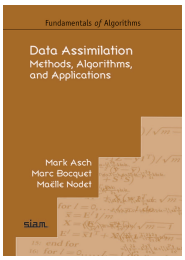
1.1 MATLAB Guide, Third Edition



By Desmond J. Higham and Nicholas J. Higham
Publisher: SIAM
ISBN: 978-1-611974-65-2, xxvi + 476 pages
Published: December 2016
http://bookstore.siam.org/OT150/

About the book: MATLAB is an interactive system for numerical computation that is widely used for teaching and research in industry and academia. It provides a modern programming language and problem solving environment, with powerful data structures, customizable graphics, and easy-to-use editing and debugging tools.

1.2 Data Assimilation: Methods, Algorithms, and Applications



By Mark Asch, Marc Bocquet, and Maëlle Nodet
Publisher: SIAM
Series: Fundamentals of Algorithms, Vol. 11
ISBN: 978-1-611974-53-9, xviii + 306 pages
Published: December 2016
http://bookstore.siam.org/FA11/

About the book: Data assimilation is an approach that combines observations and model output, with the objective of improving the latter. This book places data assimilation into the broader context of inverse problems and the theory, methods, and algorithms that are used for their solution. It provides a framework for, and insight into, the inverse problem nature of data assimilation, emphasizing "why" and not just "how." Methods and diagnostics are emphasized, enabling readers to readily apply them to their own field of study.

Chair's Column

This will be my final column as your Chair and I wanted to recap some of our recent accomplishments. SIAG/OPT continues to be the third largest SIAG within SIAM with 1172 members as of last year. I was happy to see our student numbers rebound from 2014 and we have added 189 new student members who now account for approximately 49% of our membership.

In the last column we talked about the upcoming elections and I'm happy to report who our new officers will be. Tamás Terlaky was elected as the new Chair and Andreas Waechter will be joining him as the new Vice Chair. Michael Friedlander will return as the Program Director for the SIAG. Finally, James Luedtke was elected as the SIAG Secretary/Treasurer. I want to congratulate all four on their election and I hope they will find their positions as rewarding as I have. I also wanted to thank everybody who agreed to run for office, all of whom deserve our deep gratitude.

Planning continues for our triennial conference, which will be held in Vancouver, British Columbia on May 22–25, 2017. From the looks of it, it is shaping up to be one of the biggest and best conferences ever. We had 130 minisymposia submitted, which is 12 more than in OP14 and in total we will have over 156 sessions total. We are in the final stages of scheduling the program and by the time you read this column, it will hopefully be out. I'm really looking forward to being there.

As a quick reminder, the conference will feature two minitutorials, both of which I encourage you to check out. The first is on Stochastic Optimization for Machine Learning, organized by Francis Bach and Mark Schmidt. The second will be on Optimal Power Flow and is being organized by Alper Atamturk and Daniel Bienstock.

Finally, I would like to add my sincerest thanks to the entire SIAG/OPT membership for allowing me the great privilege of serving as your Chair. It has been an incredible three years and I have enjoyed every minute of it.

See you in Vancouver!

Juan Meza, SIAG/OPT Chair
School of Natural Sciences, University of California, Merced, Merced, CA 95343, USA, jcmeza@ucmerced.edu, http://bit.ly/1G8HUxO