Internat. Math. Nachrichten
Nr. 228 (2015), 5–15

A Personal Introduction to Theoretical Dictionary Learning

Karin Schnass
University of Innsbruck

When I was asked to write an introduction to my research area dictionary learning I was excited and said yes. Then I remembered that there is already a very readable review paper doing exactly that, [41]. Since I could not do it better I decided to do it differently.

1 Sparsity and Dictionaries

I started to get interested in dictionary learning in 2007 at the end of my 2nd PhD year. My PhD topic was roughly sparsity and dictionaries, as this was what Pierre (Vandergheynst), my advisor, made almost all the group do to some degree. Since the group was a happy mix of computer scientists, electric engineers and mathematicians led by a theoretical physicist, a dictionary Φ was defined as a collection of K unit norm vectors $\phi_k \in \mathbb{R}^d$ called atoms. The atoms were stacked as columns in a matrix, which by abuse of notation was also referred to as the dictionary, that is $\Phi = (\phi_1, \dots, \phi_K) \in \mathbb{R}^{d \times K}$. A signal $y \in R^d$ was called sparse in a dictionary Φ if up to a small approximation error or noise it could be represented as linear combination of a small (sparse) number of dictionary atoms,

$$y = \sum_{k \in I} \phi_k x_k + \eta = \Phi_I x_I + \eta \quad \text{or} \quad y = \Phi x + \eta \quad \text{with} \quad \|x\|_0 = |I| = S, \tag{1}$$

where $\|\cdot\|_0$ counts the non zero components of a vector or matrix. The index set I storing the non zero entries was called the support with the understanding that for the sparsity level $S = |I|$ we have $S \ll d \leq K$ and that $\|\eta\|_2 \ll \|y\|_2$ or even better $\eta = 0$. Complications like infinite dimensions were better left alone since already the finite dimensional setting led to enough problems. The foremost problem being that as soon as the number of atoms exceeds the dimension, $K > d$, finding the best S-sparse approximation to a signal becomes NP-hard, [8]. And

ISSN 0020-7926

while having an S-sparse approximations is useful for storing signals - store S values and S addresses instead of d values - or for denoising signals - throw away η, looking through $\binom{K}{S}$ possible index sets to find this best S-sparse approximation is certainly not practical. Thus people were using suboptimal but faster approximation routines and the pet routines used in the group were (Orthogonal) Matching Pursuit, [32, 37, 9] and the Basis Pursuit Principle, [15, 11]. Matching Pursuits are greedy algorithms, which iteratively try to construct a best S-term approximation. So given a signal y, initialise $a = 0$, $r = y$, $I = \emptyset$ and then for S steps do:

- Find $i = \operatorname{argmax}_k |\langle r, \phi_k \rangle|$.
- Update the support, the approximation and the residual as

$$\begin{aligned} &I = I \cup i, \\ &a = a + \langle r, \phi_k \rangle \phi_k \text{ (MP)} \quad \text{resp.} \quad a = \Phi_I \Phi_I^\dagger y \text{ (OMP)} \\ &r = y - a. \end{aligned}$$

The Basis Pursuit Principle (BP) on the other hand is a convex relaxation technique. Assuming that an S-sparse representation of y exists, instead of solving the non-convex optimisation problem

$$(P_0) \qquad \min \|x\|_0 \quad \text{s.t.} \quad y = \Phi x \tag{2}$$

one solves the relaxed, convex problem

$$(P_1) \qquad \min \|x\|_1 \quad \text{s.t.} \quad y = \Phi x \tag{3}$$

and hopes that the solutions coincide. The relaxed problem further has the advantage that even if y is contaminated by noise the solution $\hat{x}$ will be sparse in the sense that its S largest (in absolute) components will provide a good sparse approximation to y. The big questions were, when is an S-sparse representation/approximation unique and under which conditions would a suboptimal routine be able to recover it. Since for a dictionary being an orthogonal basis the answer to both problems was 'always', the first answers for a more general overcomplete dictionary (with $K > d$) were based on the concept that the dictionary was almost orthogonal. So if the largest inner product between two different atoms, called coherence $\mu = \max_{k,j:k \neq j} |\langle \phi_k, \phi_j \rangle|$, was small, sparse representations with $S \leq (2\mu)^{-1}$ were shown to be unique and both greedy and convex relaxation methods would work, [49]. For a flavour of how things look like in more general, infinite settings a good starting point is [47]. Unfortunately this bound meant that in order for the best sparse approximation to be recoverable the dictionary had to be very incoherent (this incoherence being limited by the Welch bound according to which $\mu^2 \geq \frac{K-d}{d(K-1)} \approx \frac{1}{d}$) or the signal had to be very sparse, $S \lesssim \sqrt{d}$. Since in practice both schemes seemed to work fine for relatively coherent dictionaries resp. much larger sparsity levels, the coherence bound for sparse recovery was generally regarded as pessimistic and people (by now including me) were hunting for ways to go around it, [20, 49, 45]. One breakthrough was in 2006, when J. Tropp could show that on average BP would be successful for sparsity levels $S \lesssim \mu^{-2}$, [50]. Following the rule that what works for BP should also work for (O)MP, I tried to prove the analogue result for (O)MP, failing horribly, but at least coming up with average case results for thresholding, [44], and together with Rémi Gribonval, Holger Rauhut and Pierre Vandergheynst average case results for multichannel OMP, [21].

However 2006 was foremost the year when compressed sensing started to be all the rage, [10, 7], with the restricted isometry property (RIP) being undoubtedly one of the most elegant ways to go around the coherence bound. A compressed sensing matrix Φ is said to have the RIP with isometry constant δ_S if for all subsets I of size S and all coefficient sequences x one has,

$$(1-\delta_S)\|x\|_2^2 \leq \|\Phi_I x\|_2^2 \leq (1+\delta_S)\|x\|_2^2, \tag{4}$$

or in other words if the spectrum of $\Phi_I^\star \Phi_I$ is included in $[1-\delta_S, 1+\delta_S]$. Note that contrary to a dictionary a compressed sensing matrix does not need to have normalised columns but if it has the RIP its column norms will be bounded by $\sqrt{1 \pm \delta_S}$. The RIP turned out to be the magic ingredient, based on which one could prove that both greedy methods, [35, 36], and convex relaxation schemes, [7], could recover a sparse representation, that is x from $y = \Phi x$, and which was possible to have as long as $S \log S \lesssim d$. The drawback was that the only matrices one could be reasonably sure to have the RIP property in the regime $S \log S \approx d$, where based on random constructions. For a deterministic matrix the only feasible way to ensure it having RIP was to use a coherence based bound such as $\delta_S \leq (S-1)\mu$, which brought you back to square number one, [48]. Still almost everybody who had been doing sparse approximation before happily turned to the investigation of compressed sensing, such as extension to signals sparse in a general (orthogonal) basis, that is recover Bx from $y = \Phi Bx$ for a given basis B, design of matrices with RIP or recovery algorithms. And in line with the trend Holger, Pierre and me had a look at how compressed sensing would work for signals that are sparse in an overcomplete dictionary, [40]. We also tried but failed to prove that OMP would work if the sensing matrix had the RIP, as it later turned out with good reason, [39].

Still after 1.5 years of working on sparse recovery, compressed sensing, where you were free to choose the dictionary/sensing matrix, seemed like cheating. And weren't people forgetting that in order for compressed sensing to be applicable you first needed a dictionary to provide sparse representations? So I started to get interested in dictionary learning.

2 Dictionary Learning

The goal of dictionary learning is to find a dictionary that will sparsely represent a class of signals, meaning given a set of signals y_n, which honouring the tradition are stored as columns in a matrix $Y = (y_1 \dots y_N)$, we want to find a dictionary Φ and a coefficient matrix $X = (x_1 \dots x_N)$ such that

$$Y = \Phi X \quad \text{and} \quad X \text{ sparse.} \tag{5}$$

The 1996 paper by Olshausen and Field, [13], where a dictionary is learned on patches of natural images, is widely regarded as the mother contribution to dictionary learning, but of course since I had started my PhD in 2005 I was ignorant of most things having happened before 2004, [12, 31, 30, 29], and so the first dictionary learning algorithm I encountered was K-SVD in 2006, [3]. K-SVD is an alternating minimisation algorithm, which tries to solve the problem

$$(Q_2) \qquad \min \|Y - \Phi X\|_F \quad \text{s.t.} \quad \|x_n\|_0 \leq S \quad \text{and} \quad \Phi \in \mathcal{D}, \tag{6}$$

where $\mathcal{D}$ is defined as $\mathcal{D} = \{\Phi = (\phi_1, \dots, \phi_K) : \|\phi_k\|_2 = 1\}$, by alternating between sparsely approximating the signals in the current version of the dictionary and updating the dictionary. In particular given the current version of the dictionary Ψ and training signals y_n in each iteration it does the following.

- For all n try to find $\min \|y_n - \Phi x_n\|_2$ such that $\|x_n\|_0 \leq S$ using (O)MP/BP to update X.
- For all k construct the residual matrix E_k, by concatenating as its columns the residuals $r_n = y_n - \sum_{j \neq k} x_n(j)\psi_j$ for all n where $x_n(k) \neq 0$.
- Update the k-th atom to be the left singular vector associated to the largest singular value of E_k. (Optionally update the coefficients $x_n(k) \neq 0$ in X.)

K-SVD worked like a charm for all sensible set-ups I could imagine and all signal sizes my computer could handle. Still I thought that there should be a simpler way. Unfortunately all my efforts to explain to MATLAB how to find dictionaries in a simpler way failed and beginning of 2007 I asked Pierre for permission to finish the PhD early on grounds of 'never ever being able to find anything useful again'. The request was denied with a motivating speech and a 'Karin, go back to your office' and a couple of weeks later I was on my way to Rémi Gribonval in Rennes, which in March felt a lot like Siberia.

Also Rémi had started to become interested in dictionary learning and since K-SVD seemed like the greedy Matching Pursuit type of way, the obvious thing to do was to try the Basis Pursuit way. So starting with the naive (unstable, intractable, nightmarish) optimisation problem

$$(R_0) \qquad \min \|X\|_0 \quad \text{s.t.} \quad Y = \Phi X \quad \text{and} \quad \Phi \in \mathcal{D}, \tag{7}$$

we replaced the zeros with ones to get

$$(R_1) \qquad \min \|X\|_1 \quad \text{s.t.} \quad Y = \Phi X \quad \text{and} \quad \Phi \in \mathcal{D}, \tag{8}$$

where $\|X\|_1 := \sum_n \|x_n\|_1$, to get a stable but unfortunately not a convex optimisation problem. Indeed while (R_1) is definitely the more tractable problem, since the objective function is continuous and piecewise linear, unlike (P_1) it is not convex because the constraint manifold $\mathcal{D}$ is not convex. Also it is easy to see that the problem is invariant under sign changes and permutations of the dictionary atoms, meaning that for every local minimum there are $2^K K! - 1$ other equivalent local minima.

Faced with our creation, Rémi and I asked ourselves, shall we implement it or analyse it? We decided to analyse it because it seemed easier. The question we wanted to answer was the following: Assume that we have a dictionary Φ_0 and sparse coefficients X_0. Under which conditions is there a local minimum of (R_1) at the pair (Φ_0, X_0) or in other words when can you identify the dictionary as local minimum of (R_1)? The next few days we spent in the seminar room, Rémi calculating the tangent space to the constraint manifold and finding a first order formulation of the problem and me pointing out in detail where it would go wrong and how it was hopeless in general. I left Rennes after two weeks and spent the rest of spring and summer going on an unreasonable amount of holidays, forgetting all about dictionary learning.

However, in autumn Rémi sent Pierre and me an email with a paper draft called dictionary identifiability, that contained geometric conditions when a basis-coefficient pair would constitute a local minimum of (R_1) and our three names on it. Pierre responded fast as lightening saying, I did not contribute at all, I should not be on the paper, and after being honest with myself I tried to do the same thing, but Rémi replied, no, I'd like to keep you on it, and submitted it to a conference, [23].

Reading the paper I learned that dictionary learning was also called sparse coding, that the ℓ_1 approach was not new, [52, 38], and that the field of dictionary identification had another origin in the blind source separation community. There the dictionary atoms were called the sources, the coefficients the mixing matrix and the signals the mixtures. Also one of the first theoretical insights into dictionary identification, that is, how well can you identify your sources from the mixtures, apparently came from this community, [18, 4].

The part of the paper I liked most was the short sketch how to turn the rather technical, deterministic result into a simple result by assuming that the training signals were following a random sparse model. Feeling that I still owed my contribution, I set myself to work and we started digging through concentration of measure results and assembling them to make the sketch precise, succeeding first for orthogonal bases, [22], and finally for general bases, [24]. We did not succeed in extending our probabilistic analysis to overcomplete dictionaries, but managed to write a summary of all our results, [25], so in March 2009 I could defend my thesis including a Chapter 7 on dictionary identification.

After that I did not think about dictionary identification for a while, first because at my new postdoc job Massimo (Fornasier) was paying me to think about identifying ridge functions from point queries, [14], second because my daughter was highly objective to the idea that mum would spend any time away from home not being her personal slave and third because I was again trying to prove average case results for OMP, using decaying coefficients as additional assumption. This time I failed slightly more gracefully in the sense that in some highly unrealistic setting there would have been a result.

Still failing paid off because end of 2010 I had an idea about dictionary learning based on the decaying coefficient assumption. If all atoms in the dictionary were equally likely to give the strongest contribution to a signal, then a very simple way to recover the dictionary should be through the maximisation program

$$(\tilde{Q}_2) \qquad \max_{\Phi} \sum_n \|\Phi^\star y_n\|_\infty^2. \tag{9}$$

Following the approach that had proven successful for (R_1) I started to find the first order formulation of the problem based on the tangent space and found out that if there was a maximum at the original dictionary, it had to be a second order maximum. As the sophisticated method had failed I resorted to a brute force attack, assuming that the signals were generated from an orthonormal basis and decaying coefficients with random signs, and got a first result. Very excited, I told Rémi at the SMALL workshop about the idea and after listening patiently he said, mmmh that sounds a lot like K-SVD. Had I reinvented the wheel? No I had a first theoretical result, showing that wheels rolled.

However, this was not the most exciting part of the SMALL workshop. The most exciting part was John Wright's talk on how to extend ℓ_1-based dictionary identification to overcomplete dictionaries, [17], and his personal confirmation that implementation of a descent algorithm was hard (in somewhat more colourful words), [16]. I was motivated to do dictionary learning again and since it was also time to look for a new job, I invented a project on dictionary learning, and then hoped for one year that someone would agree to fund it. In the meantime I followed a higher calling as personal slave to now two children.

Thus I missed the development of an interesting line of results on the sample complexity of dictionary learning, [33, 51, 34, 19]. These results characterise how well the sparse approximation performance of a dictionary (for example a learned one but also a designed one) on a finite sample of training data extrapolates to future data. I also missed the development of ER-SpUD, the first algorithm which could be proven to globally recover a basis, [46], and the extension of ℓ_1-based local dictionary identification to noisy data, [26, 27].

Luckily in May 2012 the project was accepted, so I could not only continue to analyse $(\tilde{Q}_2)$ but also extend it and uncover the relation to K-SVD or rather to (Q_2), so that in early 2013 I had theoretical results indicating why K-SVD worked,

[43]. However, the most interesting development of 2013 was that two research groups independently derived algorithms, that could be proven to globally recover an overcomplete dictionary from random sparse signals, [6, 2]. Their similar approach was based on finding overlapping clusters, each corresponding to one atom, in a graph derived from the correlation matrix $Y^\star Y$ and as such radically different from the optimisation based approaches, that had led to all previous results. One group then proved local convergence properties for an alternating minimisation algorithm, which alternates between sparsely approximating the signals in the current version of the dictionary and updating the dictionary, similar to K-SVD, [1], while the other group tried to break the coherence barrier, [5].

Indeed all results mentioned so far were valid at best for sparsity levels $S \leq O(\mu^{-1})$ or under a RIP-condition on the dictionary to be recovered, meaning under the same conditions that sparse recovery was guaranteed to work. This was somewhat frustrating in view of the fact that both BP or thresholding would on average work well for sparsity levels $S \leq O(\mu^{-2})$, [50, 44] and that in dictionary learning one usually faces a lot of average signals. So I was quite proud to scratch the coherence barrier by showing that locally dictionary identification is stable for sparsity levels up to $S \leq O(\mu^{-2})$, [42].

Now at the beginning of 2015, looking at the handful of dictionary identification results so far, it is interesting to see the two origins - sparse approximation and blind source separation - represented by the two types of results, based on optimisation on one hand and on graph clustering algorithms on the other hand. Comparing the quality of the results in terms of sample complexity, sparsity level and computational complexity is difficult, see [27] for a good attempt, as they all rely on different signal models, and it is hard to entangle the strength of the signal model - (no) noise, (no) outliers, (in)exactly sparse - from the strength of the approach. One attempt at understanding the sample complexity based solely on the signal model from an information theoretic point of view can be found in [28]. Still so far graph based algorithms are the only methods with global guarantees while optimisation schemes seem to be locally more robust to noise, outliers and coherence. In short this means that I will not get bored with dictionary learning for a while as there is plenty of work to be done, for instance trying to marry globality with robustness, deriving blind learning schemes that decide the sparsity level and dictionary size for themselves or extending the results to more realistic signal models. Moreover it is high time to take another shot a average case results for OMP.

References

[1] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used overcomplete dictionaries via alternating minimization. *COLT*

2014 (arXiv:1310.7991), 2014.

[2] A. Agarwal, A. Anandkumar, and P. Netrapalli. Exact recovery of sparsely used overcomplete dictionaries. *COLT 2014 (arXiv:1309.1952)*, 2014.

[3] M. Aharon, M. Elad, and A.M. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing.*, 54(11):4311–4322, November 2006.

[4] M. Aharon, M. Elad, and A.M. Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Journal of Linear Algebra and Applications*, 416:48–67, July 2006.

[5] S. Arora, A. Bhaskara, R. Ge, and T. Ma. More algorithms for provable dictionary learning. *arXiv:1401.0579*, 2014.

[6] S. Arora, R. Ge, and A. Moitra. New algorithms for learning incoherent and overcomplete dictionaries. *COLT 2014 (arXiv:1308.6273)*, 2014.

[7] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[8] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive Approximation*, 13:57–98, 1997. Springer-Verlag New York Inc.

[9] G. M. Davis, S. Mallat, and Z. Zhang. Adaptive time-frequency decompositions with matching pursuits. *SPIE Optical Engineering*, 33(7):2183–2191, July 1994.

[10] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[11] D.L. Donoho and M. Elad. Maximal sparsity representation via ℓ^1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, March 2003.

[12] K. Engan, S.O. Aase, and J.H. Husoy. Method of optimal directions for frame design. In *ICASSP99*, volume 5, pages 2443–2446, 1999.

[13] D.J. Field and B.A. Olshausen. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[14] M. Fornasier, K. Schnass, and J. Vybíral. Learning functions of few arbitrary linear parameters in high dimensions. *Foundations of Computational Mathematics*, 2012.

[15] J. J. Fuchs. Extension of the pisarenko method to sparse linear arrays. *IEEE Transactions on Signal Processing*, 45(2413-2421), October 1997.

[16] Q. Geng, H. Wang, and J. Wright. Algorithms for exact dictionary learning by ℓ^1-minimization. *arXiv:1101.5672*, 2011.

[17] Q. Geng, H. Wang, and J. Wright. On the local correctness of ℓ^1-minimization for dictionary learning. *arXiv:1101.5672*, 2011.

[18] P. Georgiev, F.J. Theis, and A. Cichocki. Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Transactions on Neural Networks*, 16(4):992–996, 2005.

[19] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert. Sample complexity of dictionary learning and other matrix factorizations. *arXiv:1312.3790*, 2013.

[20] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12):3320–3325, December 2003.

[21] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst. Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. *Journal of Fourier Analysis and Applications*, 14(5):655–687, 2008.

[22] R. Gribonval and K. Schnass. Dictionary identifiability from few training samples. In *EUSIPCO08*, 2008.

[23] R. Gribonval and K. Schnass. Some recovery conditions for basis learning by l1-minimization. In *ISCCSP08*, March 2008.

[24] R. Gribonval and K. Schnass. Basis identification from random sparse samples. In *SPARS09*, 2009.

[25] R. Gribonval and K. Schnass. Dictionary identifiability - sparse matrix-factorisation via l_1-minimisation. *IEEE Transactions on Information Theory*, 56(7):3523–3539, July 2010.

[26] R. Jenatton, F. Bach, and R. Gribonval. Local stability and robustness of sparse dictionary learning in the presence of noise. Technical Report 00737152 (arXiv:1210.0685), INRIA - HAL, 2012.

[27] R. Jenatton, F. Bach, and R. Gribonval. Sparse and spurious: dictionary learning with noise and outliers. *arXiv:1407.5155*, 2014.

[28] A. Jung, Y. Eldar, and N. Görtz. Performance limits of dictionary learning for sparse coding. In *EUSIPCO14 (arXiv:1402.4078)*, pages 765 – 769, 2014.

[29] K. Kreutz-Delgado, J.F. Murray, B.D. Rao, K. Engan, T. Lee, and T.J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Computations*, 15(2):349–396, 2003.

[30] K. Kreutz-Delgado and B.D. Rao. FOCUSS-based dictionary learning algorithms. In *SPIE 4119*, 2000.

[31] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computations*, 12(2):337–365, 2000.

[32] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[33] A. Maurer and M. Pontil. K-dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

[34] N.A. Mehta and A.G. Gray. On the sample complexity of predictive sparse coding. *arXiv:1202.4050*, 2012.

[35] D. Needell and J.A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied Computational Harmonic Analysis*, 26(3):301–321, 2009.

[36] D. Needell and R. Vershynin. Uniform Uncertainty Principle and signal recovery via Regularized Orthogonal Matching Pursuit. *Foundations of Computational Mathematics*, DOI: 10.1007/s10208-008-9031-3.

[37] Y. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal Matching Pursuit : recursive function approximation with application to wavelet decomposition. In *Asilomar Conf. on Signals Systems and Comput.*, 1993.

[38] M.D. Plumbley. Dictionary learning for ℓ_1-exact sparse coding. In M.E. Davies, C.J. James, and S.A. Abdallah, editors, *International Conference on Independent Component Analysis and Signal Separation*, volume 4666, pages 406–413. Springer, 2007.

[39] H. Rauhut. On the impossibility of uniform sparse reconstruction using greedy methods. *Sampl. Theory Signal Image Process.*, 7(2):197–215, 2008.

[40] H. Rauhut, K. Schnass, and P. Vandergheynst. Compressed sensing and redundant dictionaries. *IEEE Transactions on Information Theory*, 54(5):2210–2219, 2008.

[41] R. Rubinstein, A. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[42] K. Schnass. Local identification of overcomplete dictionaries. *accepted to Journal of Machine Learning Research (arXiv:1401.6354)*, 2014.

[43] K. Schnass. On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD. *Applied Computational Harmonic Analysis*, 37(3):464–491, 2014.

[44] K. Schnass and P. Vandergheynst. Average performance analysis for thresholding. *IEEE Signal Processing Letters*, 14(11):828–831, 2007.

[45] K. Schnass and P. Vandergheynst. Dictionary preconditioning for greedy algorithms. *IEEE Transactions on Signal Processing*, 56(5):1994–2002, 2008.

[46] D. Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries. In *COLT 2012 (arXiv:1206.5882)*, 2012.

[47] V.N. Temlyakov. Greedy approximation. *Acta Numerica*, 17:235–409, 2008.

[48] A. Tillmann and M.E. Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2014.

[49] J.A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, October 2004.

[50] J.A. Tropp. On the conditioning of random subdictionaries. *Applied Computational Harmonic Analysis*, 25(1-24), 2008.

[51] D. Vainsencher, S. Mannor, and A.M. Bruckstein. The sample complexity of dictionary learning. *Journal of Machine Learning Research*, 12(3259-3281), 2011.

[52] M. Zibulevsky and B.A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computations*, 13(4):863–882, 2001.

Author's address: Karin Schnass, Department of Mathematics, University of Innsbruck. Technikerstraße 19a, A-6020 Innsbruck. e-mail *karin.schnass@uibk.ac.at*