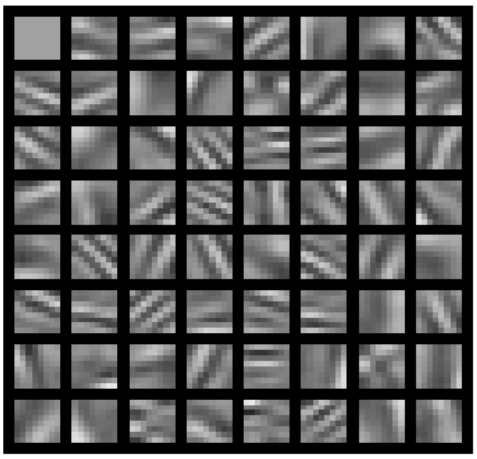
Dictionary learning

& related topics



Habilitation thesis

by

Karin Schnass

Submitted to the University of Innsbruck

Innsbruck, November 13, 2018

Chapter 1

Summary

In total the thesis contains the following articles (in order of publication date):

1. K. Schnass and J. Vybiral, *Compressed learning of high-dimensional sparse functions*, In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.3924-3927, 2011.
2. M. Fornasier, K. Schnass and J. Vybiral *Learning functions of few arbitrary linear parameters in high dimensions*, Foundations of Computational Mathematics, 12(2):229–262, 2012.
3. K. Schnass, *On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD,* Applied and Computational Harmonic Analysis, 37(3):464–491, 2014.
4. K. Schnass, *Local identification of overcomplete dictionaries,* Journal of Machine Learning Research, 16(Jun): 1211–1242, 2015.
5. K. Schnass, *A personal introduction to theoretical dictionary learning*, Internationale Mathematische Nach-richten, 228:5–15, 2015.
6. V. Naumova and K. Schnass, *Dictionary learning from incomplete data for efficient image restoration,* In 2017 25th European Signal Processing Conference (EUSIPCO), 2017.
7. V. Naumova and K. Schnass, *Fast dictionary learning from incomplete data*, EURASIP Journal on Advances in Signal Processing, 2018(12), 21 pages, 2018.
8. K. Schnass, *Convergence radius and sample complexity of ITKM algorithms for dictionary learning*, Applied and Computational Harmonic Analysis, 45(1):22–58, 2018.
9. M. Sandbichler and K. Schnass, *Online and stable learning of analysis operators*, IEEE Transactions on Signal Processing (early access), DOI:10.1109/TSP.2018.2878540, 2018.
10. K. Schnass, *Average performance of Orthogonal Matching Pursuit (OMP) for sparse approximation*, IEEE Signal Processing Letters (arXiv:1809.06684), 25(12):1865–1869, 2018.
11. K. Schnass, *Dictionary learning - from local towards global and adaptive*, arXiv:1804.07101, 49 pages, 2018.
12. F. Teixeira and K. Schnass, *Compressed dictionary learning*, arXiv:1805.00692, 23 pages, 2018.

For Article 10 the preprint version available on arXiv instead of the journal version is included. It contains more detailed proofs as well as additional information and motivations that facilitate understanding but due to the page limit could not be included in the journal version.

1.1 Outline

The main topic of this thesis is dictionary learning, in particular theoretical dictionary learning. The goal of dictionary learning is to find a compact respresentation system for a collection of signals. So, for signals $y_n \in \mathbb{R}^d$, stored as columns in a data matrix $Y = (y_1, \dots, y_N)$, one wants to find a decomposition into a dictionary, that is, a collection of K normalised vectors ϕ_k, stored as columns in the dictionary matrix $\Phi = (\phi_1, \dots, \phi_K)$, and sparse coefficients $X = (x_1, \dots, x_N)$, meaning every column x_n has only a few non-zero entries,

$$Y \approx \Phi X, \quad \text{with} \quad X \text{ sparse} \quad \text{and} \quad d \leq K \ll N.$$

The first article included, *A personal introduction to theoretical dictionary learning (Schnass)*, is not a research but a popular science article that provides an easy-to-read introduction to dictionary learning. It gives an overview over the motivations for dictionary learning, the origins of the field consisting mainly in the development of algorithms and collects all results until 2015 in the relatively young sub-field of theoretical dictionary learning. The autobiographical component further serves as background and as link to the other topics included in this thesis.

The second article, *On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD (Schnass)*, contains the first theoretical results shedding light on the performance of the famous K-SVD dictionary learning algorithm by Aharon, Elad & Bruckstein. Given a sparsity level S and a dictionary size K, K-SVD tries to solve

$$\min_{X \in \mathcal{X}_S, \Phi \in \mathcal{D}_K} \|Y - \Phi X\|_F,$$

where $\mathcal{X}_S$ is the set of matrices with at most S nonzero entries per column and $\mathcal{D}_K$ the set of dictionaries with K-atoms, by alternating between finding the best dictionary for a fixed coefficients matrix and finding the best coefficients for a fixed dictionary. The article gives a characterisation when a generating dictionary Φ_0, meaning $Y = \Phi_0 X_0$ and X_0 follows an approximately S-sparse random model, is near a local minimiser of the programme above. One of the main findings is that when the generating dictionary Φ_0 is not a tight frame or the signals are not exactly S-sparse, the local minimiser may not get arbitrarily close to the generating dictionary as the number of training signals goes to infinity.

To overcome this problem *Local identification of overcomplete dictionaries, (Schnass)*, proposes an alternative optimisation programme. It is shown that with increasing number of training signals the local minimiser of this programme indeed converges to the generating dictionary. The draw-back of the simple alternating minimisation algorithm - iterative thresholding and K (signal) means - associated to the new programme is that in experiments on synthetic data it does not converge globally. Also the derived number of training signals necessary to have a given distance between minimising and generating dictionary grows quadratically in the number of atoms.

As a next step *Convergence radius and sample complexity of ITKM algorithms for dictionary learning (Schnass)*, therefore studies the convergence behaviour of the algorithm directly. The derived sample complexity for the algorithm rather than the programme is only linear in the number of atoms and inverse quadratically in the required precision. Based on the analysis the article further introduces an adaption to the algorithm replacing signal by residual means, for which an even lower sample complexity (inverse in the required precision) can be shown. While the convergence radius proven for the modified algorithm is lower than for the simple version, in numerical experiments the residual based algorithm actually exhibits global convergence properties. Further, on image data it proves to be an interesting alternative to K-SVD, learning dictionaries of the same quality in a fraction of the time.

Fast dictionary learning from incomplete data (Naumova & Schnass) was motivated by the idea of applying dictionaries, learned with the newly available fast algorithm, for prediction on medical signals, in particular, blood glucose measurements. Unfortunately, most of the available signals proved to be incomplete, since the measurement device implanted under the skin is quite delicate and frequently produces signal drop-outs. The article therefore addresses how to learn dictionaries from partly erased signals and applies the learned dictionaries to image inpainting. *Dictionary learning from incomplete data for efficient image restoration (Naumova & Schnass)* is an extension of the previous article providing an adaptive choice of the low rank part of the dictionary and inpainting of colour images.

The computational advantages of the iterative thresholding and K residual means algorithm (ITKrM) over K-SVD are further increased in *Compressed dictionary learning (Teixeira & Schnass)*. The most costly operation in ITKrM, the matrix multiplication between the dictionary matrix and the data matrix, is replaced with an efficient approximative multiplication and the effect of the approximation on the convergence radius is analysed. The compressed version of the algorithm is then used to learn (undercomplete) dictionaries in very high dimensions - up to one quarter of a million variables - and on raw audio data.
Finally *Dictionary learning - from local towards global and adaptive* addresses the irksome fact that in *Convergence radius and sample complexity of ITKM algorithms for dictionary learning (Schnass)* the algorithm, for which the derived convergence radius is smaller, exhibits better global convergence properties in simulations. It provides relaxed conditions under which the residual based algorithm is a contraction and characterises its spurious fixed points. The characterisation is then used to derive a strategy to escape from spurious fixed points and to jump directly to the generating dictionary. The strategy, which is based on identification and (smart) replacement of coherent atoms, is then decoupled into independent pruning and adding of atoms. The resulting adaptive dictionary learning algorithm - automatic choice of the sparsity level and dictionary size - is shown to correctly identify dictionaries on synthetic data and to learn meaningful dictionaries on image data.

The second chapter of this thesis, corresponding to *Online and stable learning of analysis operators (Sandbichler & Schnass)*, addresses the dual problem to dictionary learning, known as analysis operator/dictionary learning. Given data Y as above, the goal is to find an analysis operator $A \in \mathbb{R}^{K \times d}$ such that

$$AY = X \quad \text{with} \quad X \text{ sparse} \quad \text{and} \quad d \leq K \ll N.$$

The article introduces four learning algorithms (one of them sequential), characterises their stationary points and demonstrates their effectiveness both on synthetic and image data.

The last chapter contains miscelleanous results in sparse identification. *Learning functions of few arbitrary linear parameters in high dimensions (Fornasier, Schnass & Vybiral)* derives a strategy to learn functions defined on a ball around the origin in $\mathbb{R}^d$, which only depend on a small number of linear parameters, meaning $f(x) = g(Ax)$ for a $k \times d$ matrix A. Based on smoothness and variation properties of g, it characterises the number of point queries $f(x_i)$ for $i = 1 \dots m$, which are with high probability sufficient for a given precision ε of the approximation $\tilde{f}(x) = \tilde{g}(\tilde{A}x)$, meaning $\|f - \tilde{f}\|_\infty \leq \varepsilon$. This result shows when the curse of dimensionality, which normally leads to an exponential growth of the necessary number of queries ε^{-d} even for C^∞-functions, can be broken.
Compressed learning of high-dimensional sparse functions (Schnass & Vybiral) provides a radically simplified scheme and analysis for the case where the function depends only on a few variables, meaning $f(x) = g(x_{i_1}, \dots, x_{i_k})$.
The final paper included in this thesis, *Average performance of Orthogonal Matching Pursuit (OMP) for sparse approximation (Schnass)*, is the happy ending to a 10-year-struggle. It analyses the average sparse identification performance of (OMP) for sparse signals, whose coefficients have random signs and exhibit decay. The result rehabilitates greedy methods like OMP as not only cheap but potentially better alternative to convex relaxation methods such as Basis Pursuit (BP), by showing that neither OMP nor its famous competitor BP are better but that each outperforms the other in different regimes.

Acknowledgements

This thesis would not exist without the FWF (Austrian Science Fund), that has been paying my salary since the end of my PhD. It would not exist without Massimo *sensei* Fornasier, who hired me as a very part-time postdoc and on the subject of a 10-minute presentation for the START-project told me 'if you cannot explain it with pictures, you cannot explain it; blow them away!'. It would not exist either without Maria *project fairy* Mateescu, whose proof-reading of my project proposals has a 100% acceptance rate. It might have been submitted somewhere else without Doris *guardian angel* Mangott, my endless source of counsel and coffee. It would probably stop after two thirds without my co-authors and colleagues from Linz, Alghero and Innsbruck and contain much lighter results without friends &family all over the world, who have been keeping me sane throughout these years.

Thank you!