Large-scale structure-adaptive signal and image recovery

1 Introduction

Our focus is on some problems of high-dimensional statistics where the data has a simple local structure which is unknown to the statistician. This is the case, for instance, in textured image recovery, speech segmentation, and sparse recovery problems in statistical signal processing. When the structural parameters are known *a priori*, linear recovery procedures constitute a simple to use yet powerful tool of signal recovery and prediction. We propose a new generation of denoising procedures: nonlinear adaptive estimators capable of "extracting" unknown structural information from observations and perform nearly as good as the best linear estimators, and in some sense nearly as good as it is theoretically possible [1, 2].

When reduced to its simplest expression, our problem is to recover a sampled signal $x = (x_t)$, $-2n \le t \le 2n$, possibly in several dimensions, from noisy observations y; in reality, x may represent a digital sound or image file. For example, one often deals with observations

$$y_t = x_t + \sigma\xi_t$$

corrupted by the random Gaussian noise $\xi_t \sim \mathcal{N}(0,1)$. It is widely known that as long as the signal to be recovered has some inherent *local structure*, recovery can be done via *linear filtering*:

$$\widehat{x}^o_t = [\varphi^o * y]_t := \sum_{|\tau|\le n} \varphi^o_\tau\, y_{t-\tau}, \quad |t| \le n,$$

where the vector $\varphi^o = (\varphi^o_{-n}, ..., \varphi^o_n)$ is called *a filter*, and $*$ is the convolution operator. In the classical setting, the structure is supposed to be known, and hence a filter φ^o with a good statistical performance can be computed in advance, before observing the data. In practice, however, this assumption is often too restrictive. For example, adjacent pixels in images are generally well-correlated, but this correlation unpredictably breaks at the object boundaries. Hence, we are interested in a more realistic scenario, where the structure is *unknown*, so that a well-performing linear filter cannot be computed beforehand; instead, we only make a weak assumption that such a filter, an *oracle*, exists in the nature. We then ask a natural question:

> *Can one* ***adapt*** *to the unknown structure, i.e. build a data-driven recovery $\widehat{x}$ which is nearly as good as $\widehat{x}^o$, the recovery corresponding to the unknown oracle?*

2 Adaptive recovery procedures

Our developments provide a positive answer to the outlined question [3, 4]. Specifically, we suggest to use a (non-linear) recovery of the form $\widehat{x} = \widehat{\varphi} * y$ where the filter $\widehat{\varphi} = \widehat{\varphi}(y)$ is an optimal solution to some well-structured convex program such as the following one:

$$\min_{\varphi} \left\| [y - \varphi * y]^n_{-n} \right\|_2^2 + \lambda \|\mathcal{F}\varphi\|_1, \tag{1}$$

where $\|\cdot\|_p$ is the ℓ_p norm, $\mathcal{F}$ is the Discrete Fourier transform operator, and λ is the regularization parameter. In other words, one is searching for a filter which reproduces the data well, and at the same time has a low complexity as measured by the ℓ_1 norm of its Fourier spectrum.

Statistical guarantees. Earlier works [5, 6, 7] indicate that the statistical performance of adaptive recoveries corresponding to (1) and similar procedures can nearly match that of $\widehat{x}^o$ whenever there exists an oracle φ^o of a small norm which reproduces the signal with a small error. Recently, we proposed [4] a simple sufficient condition for the existence of such φ^o:

approximate shift-invariance (ASI) of the unknown signal. Informally, x is called s-ASI if it is sufficiently close to some shift-invariant subspace of dimension $s \leq n$. Several important statistical models satisfy this assumption, most notably nonparametric kernel regression and line spectral signal estimation [3]. In [4] it was shown that s-ASI implies the existence of a linear oracle φ^o with quadratic risk $\mathcal{O}(s\sigma^2/n)$, and on the other hand, the risk of the adaptive recovery (1) is $\mathcal{O}(s+\log n)$ times greater. In other words, the adaptive recovery performs nearly as well as the oracle one whenever $s \ll n$.

Numerical algorithms. The optimization problem (1) is a well-structured second-order conic program. The classical methods of choice to solve such problems are the interior point methods since they achieve a very high numerical accuracy in just a few iterations. However, interior point methods quickly become inapplicable for (1) as the sample size n grows, since in each iteration it is required to solve a dense linear system with $\mathcal{O}(n)$ variables.

Instead, we propose to use *first-order methods* to tackle (1) and similar optimization problems [8, 9, 10, 11]. The methods make use of only the first-order information about the objective, and are beneficial for (1) due to a very low iteration complexity: in each iteration they require nothing more than several elementwise vector operations and gradient computations. Moreover, for the problem at hand, the gradient computation can be expressed as a matrix-vector multiplication with a structured matrix with computational complexity quasi-linear in n. After some fine-tuning, one typically has to perform an order of $10^1 - 10^2$ iterations to attain a sufficiently low numerical error.

Applications. Let us describe some application scenarios for our techniques.

Denoising of signals with line spectra. In speech and music processing, signal can be modelled as a combination of a few *harmonics*, having a Fourier spectrum composed of several spikes. In practice, one has a limited number of noisy samples of such a signal, and would like to get rid of the additive noise, e.g. the recording or the coding losses, as well as the spectral blurring caused by the limited observation time. Similar problems arise in medical imaging, spectroscopy, and high-energy physics, with the time and spectral domains reversed: the spikes correspond to instantaneous events, and the data are noisy low-frequency measurements of the signal.

More formally, the spectrum of the signal to recover, given infinite observation time, can be modeled as the sum of spikes positioned on $0 \leq f_1, ..., f_s \leq \mathtt{fs}$, where $\mathtt{fs}$ is the sampling rate in Hz. The signal structure is described by the frequencies $f_1, ..., f_s$. Note that even without the measurement noise, because of the spectral blurring the frequencies cannot be reliably identified unless they are separated by $\mathcal{O}(\mathtt{fs}/n)$, where n is the number of observations, see e.g. [12]. Existing methods for signal denoising in this scenario essentially treat recovery in the time and the frequency domain as two equivalent problems. As a result, the performance of these methods dramatically degrades if the separation assumption is relaxed [13, 14]. Our theory of adaptive recovery manages to bridge this gap [3]: the proposed recovery procedures demonstrate a good empirical performance, and are nearly theoretically optimal without any separation assumption.

Dimension reduction. In some applications, e.g. in econometrics, one sometimes deals with *index models*, where a high-dimensional space X_{high} of explanatory variables is first linearly mapped to a lower-dimensional space X_{low}, and then the response is given by some smooth function on X_{low}; the signal structure is formalized as unknown linear mapping $X_{\text{high}} \to X_{\text{low}}$. The simplest case is the so called *single-index* [4], where the dimension of X_{low} is just one, i.e. the response changes only along one direction. Essentially, if one knew the direction, one could just average orthogonally to it, obtaining a one-dimensional regression problem with noise suppressed by averaging. Adaptive filtering procedures demonstrate good results on preliminary experiments for that application as well [4].

3 Future research directions

Inverse problems. Denoising problem described above is naturally generalized to the case of indirect observations:

$$y = Ax + \sigma\xi,$$

where A is a time-invariant linear operator corresponding to the convolution with a given filter, or *blurring*. Compared to the case of direct observations, statistical assumptions of existing methods are often too restrictive [15], and even some of the basic questions turn out to be beyond their reach (for example, the minimax rate of recovering a signal with a line spectrum from indirect observations is unknown to us). The situation becomes even more interesting in the case where the observation operator A is unknown, so called *blind deconvolution*, where the existing techniques are either non-convex and hence lack global convergence guarantees [16], or take too strong assumptions about the underlying data [17]. As such, it would be exciting to extend our techniques to this more general setting.

Geometric statistical signal processing. Many signal processing problems involve data on non-Euclidean domains, such as Riemannian manifolds or graphs. For instance, in computer graphics and vision, 3D objects are modeled as manifolds endowed with properties such as color or texture, or alternatively, as graphs arising as triangulations of these manifolds. Other relevant examples include the models of social networks [18], gene expression data, and dynamic models in neuroscience [19], in all of which one has to deal with multiple time-varying processes in the nodes of a large graph, the edges of which govern the correlation between the processes. Exploiting the underlying low-rank structure is often vital in these applications, and our techniques, after a proper generalization, can be capable of inferring this structure.

Post selection inference. Linear regression is a simple and powerful statistical technique. Not only it allows to estimate the impacts of explanatory variables in the form of regression coefficients, but it also provides confidence intervals for these estimates. However, in modern datasets, the number of candidate variables is often much larger than the sample size, whereas only a small number of them are actually relevant. In these conditions, one would prefer first to select only (supposedly) relevant variables by means of some model selection procedure, and then to regress only on these variables. The problem with this approach is that the usual confidence intervals tend to be too narrow since the inference is now performed on a model which depends on the data and may prove to be wrong with non-vanishing probability. Current quantitative explanations of this phenomenon, see e.g. [20] and [21], require some stringent assumptions and lack non-asymptotic results, so a lot can be done in this area of research.

References

[1] I. Ibragimov and R. Khasminskii, "Nonparametric estimation of the value of a linear functional in Gaussian white noise," *Theor. Probab. & Appl.*, vol. 29, pp. 1–32, 1984.

[2] D. Donoho, "Statistical estimation and optimal recovery," *Ann. Statist.*, vol. 22, pp. 238–270, 03 1994.

[3] Z. Harchaoui, A. Juditsky, A. Nemirovski, and D. Ostrovsky, "Adaptive recovery of signals by convex optimization," in *Proceedings of The 28th Conference on Learning Theory*, pp. 929–955, 2015.

[4] D. Ostrovsky, Z. Harchaoui, A. Juditsky, and A. Nemirovski, "Structure-blind signal recovery," *arXiv preprint arXiv:1607.05712*, 2016.

[5] A. Juditsky and A. Nemirovski, "Nonparametric denoising of signals with unknown local structure, i: Oracle inequalities," *Applied and Computational Harmonic Analysis*, vol. 27, no. 2, pp. 157–179, 2009.

[6] A. Juditsky and A. Nemirovski, "Nonparametric denoising of signals with unknown local structure, ii: Nonparametric function recovery," *Applied and Computational Harmonic Analysis*, vol. 29, no. 3, pp. 354–367, 2010.

[7] A. Juditsky and A. Nemirovski, "On detecting harmonic oscillations," *Bernoulli*, vol. 21, no. 2, pp. 1134–1165, 2015.

[8] A. Juditsky and A. Nemirovski, "First order methods for nonsmooth convex large-scale optimization, i: general purpose methods," *Optimization for Machine Learning*, pp. 121–148, 2011.

[9] A. Juditsky and A. Nemirovski, "First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure," *Optimization for Machine Learning*, pp. 149–183, 2011.

[10] Y. Nesterov and A. Nemirovski, "On first-order algorithms for ℓ_1/nuclear norm minimization," *Acta Numerica*, vol. 22, pp. 509–575, 5 2013.

[11] S. Sra, S. Nowozin, and S. Wright, *Optimization for machine learning*. MIT Press, 2012.

[12] M. Duarte and R. Baraniuk, "Spectral compressive sensing," *Applied and Computational Harmonic Analysis*, vol. 35, no. 1, pp. 111–129, 2013.

[13] B. Bhaskar, G. Tang, and B. Recht, "Atomic norm denoising with applications to line spectral estimation," *Signal Processing, IEEE Transactions on*, vol. 61, no. 23, pp. 5987–5999, 2013.

[14] G. Tang, B. Bhaskar, and B. Recht, "Near minimax line spectral estimation," in *Information Sciences and Systems (CISS), 2013 47th Annual Conference on*, 2013.

[15] C. Butucea and F. Comte, "Adaptive estimation of linear functionals in the convolution model and applications," *Bernoulli*, vol. 15, pp. 69–98, 02 2009.

[16] A. Bronstein, M. Bronstein, and M. Zibulevsky, "Blind source separation using block-coordinate relative Newton method," *Signal Processing*, vol. 84, no. 8, pp. 1447 – 1459, 2004.

[17] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *IEEE Trans. Inf. Theor.*, vol. 60, pp. 1711–1732, Mar. 2014.

[18] D. Lazer *et al.*, "Life in the network: The coming age of computational social science," *Science*, vol. 323(5915), 2009.

[19] M. Schwemmer, A. Fairhall, S. Denéve, and E. Shea-Brown, "Constructing precisely computing networks with biophysical spiking neurons," *The Journal of Neuroscience*, vol. 35, no. 28, pp. 10112–10134, 2015.

[20] J. Lee, D. Sun, Y. Sun, and J. Taylor, "Exact post-selection inference, with application to the lasso," *Ann. Statist.*, vol. 44, pp. 907–927, 06 2016.

[21] V. Chernozhukov, C. Hansen, and M. Spindler, "Valid post-selection and post-regularization inference: An elementary, general approach," *Annual Review of Economics*, vol. 7, no. 1, pp. 649–688, 2015.