Iterative Thresholding for Demixing Structured Superpositions in High Dimensions

Mohammadreza Soltani
Iowa State University

Chinmay Hegde
Iowa State University

Abstract

We consider the demixing problem of two (or more) high-dimensional vectors from nonlinear observations when the number of such observations is far less than the ambient dimension of the underlying vectors. Specifically, we demonstrate an algorithm that stably estimate the underlying components under general *structured sparsity* assumptions on these components. Specifically, we show that for certain types of structured superposition models, our method provably recovers the components given merely $n = \mathcal{O}(s)$ samples where s denotes the number of nonzero entries in the underlying components. Moreover, our method achieves a fast (linear) convergence rate, and also exhibits fast (near-linear) per-iteration complexity for certain types of structured models. We also provide a range of simulations to illustrate the performance of the proposed algorithm.

1 Introduction

The *demixing* problem involves disentangling two (or more) high-dimensional vectors from their linear superposition [1, 2, 3, 4, 5]. In statistical learning applications involving parameter estimation, such superpositions can be used to model situations when there is some ambiguity in the parameters (e.g., the true parameters can be treated as "ground truth" + "outliers") or when there is some existing prior knowledge that the true parameter vector is a superposition of two components. Mathematically, suppose that the parameter vector is given by $\beta = \Phi\theta_1 + \Psi\theta_2$ where $\beta, \theta_1, \theta_2 \in \mathbb{R}^p$ and Φ, Ψ are orthonormal bases. If a linear observation model is assumed, then given samples $y \in \mathbb{R}^n$ and a design matrix $X \in \mathbb{R}^{n\times p}$, the goal is to recover the parameter vector β that minimizes a loss function $\mathcal{L}(X, y; \beta)$. We focus on the sample-poor regime where the dimension far exceeds the number of samples; this regime has received significant attention from the machine learning and signal processing communities in recent years [6, 7].

However, fitting the observations according to a linear model can be restrictive. One way to ease this restriction is to assume a *nonlinear* observation model:

$$y = g(X\beta) + e = g(X(\Phi\theta_1 + \Psi\theta_2)) + e, \tag{1.1}$$

where g denotes a nonlinear *link* function and e denotes observation noise. This is akin to the *Generalized Linear Model* (GLM) and *Single Index Model* (SIM) commonly used in statistics [8]. Here, the problem is to estimate w and z from the observations y with as few samples as possible.

The above estimation problem is challenging in several different aspects: (i) there is a basic identifiability of issue of obtaining θ_1 and θ_2 even with perfect knowledge of β; (ii) there is a second identifiability issue arising from the nontrivial null-space of the design matrix (since $n \ll p$); and (iii) the nonlinear nature of g, as well as the presence of noise e can further confound recovery.

Standard techniques to overcome each of these challenges are well-known. By and large, these techniques all make some type of *sparseness* assumption on the components θ_1 and θ_2 [7]; some type of *incoherence* assumption on the bases Φ and Ψ [9, 10]; some type of *restricted strong convexity*

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

(RSC) [6]; and some type of *Lipschitz* (*restricted strong smoothness* (RSS)) assumptions on the link function g [11]. See section 2 for details.

In this short paper, we demonstrate an algorithm that stably estimate the components θ_1 and θ_2 under general *structured sparsity* assumptions on these components. Structured sparsity assumptions are useful in applications where the support patterns (i.e., the coordinates of the nonzero entries) belong to certain restricted families (for example, the support is assumed to be *group-sparse* [12]). It is known that such assumptions can significantly reduce the required number of samples for estimating the parameter vectors, compared to generic sparsity assumptions [13, 14, 15].

We note that demixing approaches in high dimensions with structured sparsity assumptions have appeared before in the literature [1, 2, 16]. However, our method differs from these earlier works in a few different aspects. The majority of these methods involve solving a convex relaxation problem; in contrast, our algorithm is manifestly *non-convex*. Despite this feature, for certain types of structured superposition models our method provably recovers the components given merely $n = \mathcal{O}(s)$ samples; moreover, our methods achieve a fast (linear) convergence rate, and also exhibits fast (near-linear) per-iteration complexity for certain types of structured models. Moreover, these earlier methods have not explicitly addressed the nonlinear observation model (with the exception of [17]). We show that under certain smoothness assumptions on g, the performance of our method matches (in terms of asymptotics) the best possible sample-complexity.

2 Preliminaries

Let $\|.\|_q$ denote the ℓ_q-norm of a vector. Denote the spectral norm of the matrix X as $\|X\|$. Denote the true parameter vector, $\theta = [\theta_1^T \ \theta_2^T]^T \in \mathbb{R}^{2p}$ as the vector obtaining by stacking the true and unknown coefficient vectors, θ_1, θ_2. For simplicity of exposition, we suppose that components θ_1 and θ_2 have block sparsity with sparsity s and block size b [13] (Analogous approaches apply for other structured sparsity models.)

The problem (1.1) is inherently unidentifiable and to resolve this issue, we need to assume that the coefficient vectors θ_1, θ_2 are distinguishable from each other. This issue is characterized by a notion of incoherence of the components θ_1, θ_2 [5].

Definition 2.1. *The bases Φ and Ψ are called ε-incoherent if $\varepsilon = \sup_{\substack{\|u\|_0 \le s,\ \|v\|_0 \le s \\ \|u\|_2 = 1,\ \|v\|_2 = 1}} |\langle \Phi u, \Psi v \rangle|$.*

For the analysis of our proposed algorithm we need the following standard definition [6]:

Definition 2.2. *$f : \mathbb{R}^{2p} \to \mathbb{R}$ satisfies Structured Restricted Strong Convexity/Smoothness (SRSC/SRSS) if:*

$$m_{4s} \le \|\nabla^2_\xi f(t)\| \le M_{4s}, \ \ t \in \mathbb{R}^{2p},$$

where $\xi = supp(t_1) \cup supp(t_2)$, for all $t_i \in \mathbb{R}^{2p}$ such that t_i belongs to $(2s, b)$ block-sparse vectors for $i = 1, 2$, and m_{4s} and M_{4s} are (respectively) the SRSC and SRSS constants. Also $\nabla^2_\xi f(t)$ denotes a $4s \times 4s$ sub-matrix of the Hessian matrix $\nabla^2 f(t)$ comprised of row/column indices indexed by ξ.

Also, we assume that the derivative of the link function is strictly bounded either within a positive interval, or within a negative interval.

3 Algorithm and main theory

In this section, we describe our algorithm which we call it *Structured Demixing with Hard Thresholding* (STRUCT-DHT) and our main theory. To solve demixing problem in (1.1), we consider the minimization of a special loss function $F(t)$ following [5]:

$$\begin{aligned} \min_{t \in \mathbb{R}^{2p}} \ \ & F(t) = \frac{1}{m} \sum_{i=1}^{m} \Theta(x_i^T \Gamma t) - y_i x_i^T \Gamma t \\ & \text{s. t. } \ t \in \mathcal{D} \end{aligned} \tag{3.1}$$

where $\Theta(x) = \int_{-\infty}^{x} g(u) du$ denotes as the integral of the link function g, $\Gamma = [\Phi \ \Psi]$, x_i is the i^{th} row of the design matrix X and $\mathcal{D}$ denotes the set of length-$2p$ vectors formed by stacking a pair of

Algorithm 1 Structured Demixing with Hard Thresholding (STRUCT-DHT)

Inputs: Bases Φ and Ψ, design matrix X, link function g, observation y, sparsity s, step size η'.
Outputs: Estimates $\widehat{\beta} = \Phi\widehat{\theta_1} + \Psi\widehat{\theta_2}$, $\widehat{\theta_1}$, $\widehat{\theta_2}$
Initialization:
$\left(\beta^0, \theta_1^0, \theta_2^0\right) \leftarrow$ RANDOM INITIALIZATION
$k \leftarrow 0$
while $k \leq N$ **do**
  $t^k \leftarrow [\theta_1^k; \theta_2^k]$ {Forming constituent vector}
  $t_1^k \leftarrow \frac{1}{m}\Phi^T X^T(g(X\beta^k) - y)$
  $t_2^k \leftarrow \frac{1}{m}\Psi^T X^T(g(X\beta^k) - y)$
  $\nabla F^k \leftarrow [t_1^k; t_2^k]$ {Forming gradient}
  $\tilde{t}^k = t^k - \eta'\nabla F^k$ {Gradient update}
  $[\theta_1^k; \theta_2^k] \leftarrow \mathcal{P}_{s;s}\left(\tilde{t}^k\right)$ {Projection}
  $\beta^k \leftarrow \Phi\theta_1^k + \Psi\theta_2^k$ {Estimating $\widehat{x}$}
  $k \leftarrow k + 1$
end while
Return: $\left(\widehat{\theta_1}, \widehat{\theta_2}\right) \leftarrow \left(\theta_1^N, \theta_2^N\right)$

(s, b) block-sparse vectors. The objective function in (3.1) is motivated by the single index model in statistics; for details, see [5]. To approximately solve (3.1), we propose STRUCT-DHT which is detailed as Algorithm 1.

At a high level, STRUCT-DHT tries to minimize loss function defined in (3.1) (tailored to g) between the observed samples y and the predicted responses $X\Gamma\widehat{t}$, where $\widehat{t} = [\widehat{\theta}_1;\ \widehat{\theta}_2]$ is the estimate of the parameter vector after N iterations. The algorithm proceeds by iteratively updating the current estimate of $\widehat{t}$ based on a gradient update rule followed by (myopic) *hard thresholding* of the residual onto the set of s-sparse vectors in the span of Φ and Ψ. Here, we consider a version of DHT [5] which is applicable for the case that coefficient vectors θ_1 and θ_2 have block sparsity. For this setting, we replace the hard thresholding step, $\mathcal{P}_{s;s}$ by component-wise block-hard thresholding [13]. Specifically, $\mathcal{P}_{s;s}(\tilde{t}^k)$ projects the vector $\tilde{t}^k \in \mathbb{R}^{2p}$ onto the set of concatenated (s, b) block-sparse vectors by projecting the first and the second half of $\tilde{t}^k$ separately.

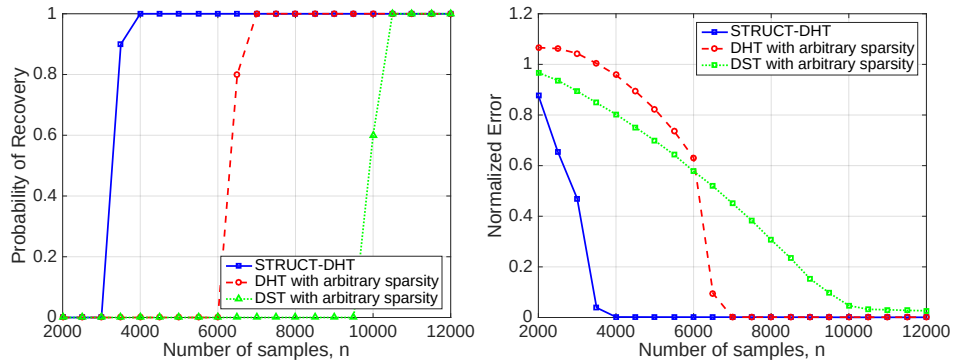
Now, we provide our main theorem supporting the convergence analysis and sample complexity (required number of observations for successful estimation of θ_1, θ_2) of STRUCT-DHT.

Theorem 3.1. *Consider the observation model* (1.1) *with all the assumption and definitions mentioned in the section 2. Suppose that the corresponding objective function F satisfies the Structured SRSS/SRSC properties with constants M_{6s} and m_{6s} such that $1 \leq \frac{M_{6s}}{m_{6s}} \leq \frac{2}{\sqrt{3}}$. Choose a step size parameter η' with $\frac{0.5}{M_{6s}} < \eta' < \frac{1.5}{m_{6s}}$. Then,* DHT *outputs a sequence of estimates (θ_1^k, θ_1^k) ($t^{k+1} = [\theta_1^k; \theta_1^k]$) such that the estimation error of the parameter vector satisfies the following upper bound (in expectation) for any $k \geq 1$:*

$$\|t^{k+1} - \theta\|_2 \leq (2q)^k \|t^0 - \theta\|_2 + C\tau\sqrt{\frac{s}{m}}, \tag{3.2}$$

where $q = 2\sqrt{1 + \eta'^2 M_{6s}^2 - 2\eta' m_{6s}}$ and $C > 0$ is a constant that depends on the step size η' and the convergence rate q. Here, θ denotes the true parameter vector defined in section 2.

Proof sketch. The proof follows the technique used to prove Theorem 4.6 in [3]. The main steps are as follows. Let $b' \in \mathbb{R}^{2p} = [b'_1; b'_2] = t^k - \eta'\nabla F(t^k)$, $b = t^k - \eta'\nabla_J F(t^k)$ where $J = \text{supp}(t^k) \cup \text{supp}(t^{k+1}) \cup \text{supp}(\theta)$ and $b'_1,\ b'_2 \in \mathbb{R}^p$ (Here, $\theta = [\theta_1; \theta_2]$ denotes the true parameter vector). Also define $t^{k+1} = \mathcal{P}_{s;s}(b') = [\mathcal{P}_s(b'_1); \mathcal{P}_s(b'_2)]$. Now, by the triangle inequality, we have: $\|t^{k+1} - \theta\|_2 \leq \|t^{k+1} - b\|_2 + \|b - \theta\|_2$. The proof is completed by showing that $\|t^{k+1} - b\|_2 \leq 2\|b - \theta\|_2$. Finally, we use the Khintchine inequality [18] to bound the expectation of the ℓ_2-norm of the restricted gradient function, $\nabla F(\theta)$ (evaluated at the true parameter vector θ) with respect to the support set J). □



(a) (b)

Figure 1: *Comparison of* DHT *with structured sparsity with other algorithms. (a) Probability of recovery in terms of normalized error. (b) Normalized error between* $\widehat{\beta} = \Phi\widehat{\theta}_1 + \Psi\widehat{\theta}_2$ *and true* β.

Inequality (3.2) indicates the linear convergence behavior of our proposed algorithm. Specifically, in the noiseless scenario to achieve κ-accuracy in estimating the parameter vector $\widehat{t} = [\widehat{\theta}_1;\ \widehat{\theta}_2]$, STRUCT-DHT only requires $\log\left(\frac{1}{\kappa}\right)$ iterations. We also have the following theorem regarding the sample complexity of Alg. 1:

Theorem 3.2. *If the rows of* X *are independent subgaussian random vectors [18], then the required number of samples for successful estimation of the components,* n *is given by* $\mathcal{O}\left(\frac{s}{b}\log\frac{p}{s}\right)$. *Furthermore, if* $b = \Omega\left(\log\frac{p}{s}\right)$, *then the sample complexity of our proposed algorithm is given by* $n = \mathcal{O}(s)$, *which is asymptotically optimal.*

Proof. The proof is similar to the proof of Theorem 4.8 in [3] where we derived upper bounds on the sample complexity by proving the RSC/RSS for the objective function F. Here, the steps are essentially the same as in [3], except that we need to compute union bound over the set of (s, b) block-sparse vectors. This set is considerably smaller than the set of *all* sparse vectors and results in an asymptotic gain in sample complexity. □

The big-Oh constant hides dependencies on various parameters, including the coherence parameter ε, as well as the upper bound and lower bounds on the derivative of the link function g.

4 Numerical results

To show the efficacy of STRUCT-DHT for demixing components with structured sparsity, we numerically compare STRUCT-DHT with ordinary DHT (which does *not* leverage structured sparsity), and also with an adaptation of a convex formulation described in [11] that we call *Demixing with Soft Thresholding* (DST). We first generate true components θ_1 and θ_2 with length $p = 2^{16}$ with nonzeros grouped in blocks with length $b = 16$ and total sparsity $s = 656$. The nonzero (active) blocks are randomly chosen from a uniform distribution over all possible blocks. We construct a design (observation) matrix following the construction of [19]. Finally, we use a (shifted) sigmoid link function given by $g(x) = \frac{1-e^{-x}}{1+e^{-x}}$ to generate the observations y. Fig 1 shows the the performance of the three algorithms with different number of samples averaged over 10 Monte Carlo trials. In Fig 1(a), we plot the probability of successful recovery, defined as the fraction of trials where the normalized error is less than 0.05. Fig 1(b) just shows the normalized estimation error for these algorithms. As we can see, STRUCT-DHT shows much better sample complexity (the required number of samples for obtaining small relative error) as compared to DHT and DST.

References

[1] M. McCoy and J. Tropp. Sharp recovery bounds for convex demixing, with applications. *Foundations of Comp. Math.*, 14(3):503–567, 2014.

[2] M. McCoy, V. Cevher, Q. Dinh, A. Asaei, and L. Baldassarre. Convexity in source separation: Models, geometry, and algorithms. *IEEE Sig. Proc. Mag.*, 31(3):87–95, 2014.

[3] M. Soltani and C. Hegde. Fast algorithms for demixing sparse signals from nonlinear observations. *arXiv preprint arXiv:1608.01234*, 2016.

[4] M. Soltani and C. Hegde. Demixing sparse signals from nonlinear observations. In *Proc. Asilomar Conf. Sig. Sys. Comp.*, Nov. 2016.

[5] M. Soltani and C. Hegde. A fast iterative algorithm for demixing sparse signals from nonlinear observations. In *Proc. IEEE Global Conf. Signal and Image Processing (GlobalSIP)*, Dec. 2016.

[6] S. Negahban, B. Yu, M. Wainwright, and P. Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Adv. Neural Inf. Proc. Sys. (NIPS)*.

[7] E. Candès. Compressive sampling. In *Proc. Int. Congress of Math.*, Madrid, Spain, Aug. 2006.

[8] S. Kakade, V. Kanade, O. Shamir, and A. Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Adv. Neural Inf. Proc. Sys. (NIPS)*, pages 927–935, 2011.

[9] M. Elad, J. Starck, P. Querre, and D. Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Appl. Comput. Harmonic Analysis*, 19(3):340–358, 2005.

[10] D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1):6–18, 2006.

[11] Z. Yang, Z. Wang, H. Liu, Y. Eldar, and T. Zhang. Sparse nonlinear regression: Parameter estimation and asymptotic inference. *arXiv preprint arXiv:1511.04514*, 2015.

[12] J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

[13] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inform. Theory*, 56(4):1982–2001, Apr. 2010.

[14] C. Hegde and R. Baraniuk. Signal recovery on incoherent manifolds. *IEEE Trans. Inform. Theory*, 58(12):7204–7214, Dec. 2012.

[15] C. Hegde, P. Indyk, and L. Schmidt. Approximation algorithms for model-based compressive sensing. *IEEE Trans. Inform. Theory*, 61(9):5129–5147, 2015.

[16] N. Rao, P. Shah, and S. Wright. Forward-backward greedy algorithms for signal demixing. In *Proc. Asilomar Conf. Sig. Sys. Comput.*, pages 437–441, 2014.

[17] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *arXiv preprint arXiv:1404.3749*, 2014.

[18] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[19] F. Krahmer and R. Ward. New and improved johnson-lindenstrauss embeddings via the restricted isometry property. *SIAM J. Math. Anal.*, 43(3):1269–1281, 2011.