Convex Multilinear Estimation and Operatorial Representations

Marco Signoretto
Katholieke Universiteit Leuven, ESAT-SCD/SISTA
Kasteelpark Arenberg 10
B-3001 Leuven (BELGIUM)
marco.signoretto@esat.kuleuven.be

Lieven De Lathauwer
Group Science, Engineering and Technology
Katholieke Universiteit Leuven, Campus Kortrijk
E. Sabbelaan 53 8500 Kortrijk (BELGIUM)
lieven.delathauwer@kuleuven-kortrijk.be

Johan A. K. Suykens
Katholieke Universiteit Leuven, ESAT-SCD/SISTA
Kasteelpark Arenberg 10
B-3001 Leuven (BELGIUM)
johan.suykens@esat.kuleuven.be

1 Introduction

In this short paper we outline a unifying framework for convex multilinear estimation, based on our recent work [15], and sketch a kernel extension to tensor-based modeling in line with [14].

Traditional tensor-based approaches often translate into challenging non-convex optimization problems that suffer from local minima. As a first contribution we consider in the next Section a general class of non-smooth convex optimization problems where a nuclear norm for tensors [9] is employed as a penalty function to enforce parsimonious solutions. For supervised learning the proposed framework allows to extend the penalized empirical risk minimization used in machine learning to develop structured (tensor-based) models. On the other hand problems like tensor completion and tensor denoising — that can be seen as unsupervised tasks — also arise as special instances of the general class of optimization problems that we consider. A common algorithm is developed to deal with these different cases. The approach builds upon existing methods for convex separable problems [10] and distributed convex optimization [5]. Furthermore being essentially a first-order scheme in the tensor unknown, the strategy we pursue can be accelerated to achieve the optimal rate of convergence in the sense of Nesterov [11]. From a methodological perspective extending the nuclear norm — and, more generally, the class of Shatten norms — from matrices to tensors [15] poses new interesting questions. For second order tensors a known result shows that the nuclear norm is the convex envelope of the rank function. For the general $N-$th order case answering whether the convex relaxation obtained with the new penalty is tight with respect to related rank-constrained formulations is an important question that goes beyond mere mathematical interest. In fact, a better understanding of these aspects might lead to more accurate convex heuristics for non-convex tensor-based problems.

Beyond non-convexity an important drawback of traditional tensor-based techniques consists of the linearity of models with respect to the data, a fact that often translates into limited discriminative power. By contrast, in the last two decades kernel models proved to be very accurate thanks to their flexibility. In Section 3 we sketch a possible approach to extend the classical tensor-based framework [14] and highlight the difference with seemingly similar ideas [17]. Whereas application of kernel methods would normally prescribe to flatten the various dimensions first, our proposal consists of mapping tensors based upon the SVD decomposition (and higher order versions thereof [6]) so that the structural information embodied in the original representation is retained.

In the following we denote scalars by lower-case letters ($a, b, c, \ldots$), vectors as capitals ($A, B, C, \ldots$) and matrices as bold-face capitals ($\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \ldots$). Tensors are written as calligraphic letters ($\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots$). We write a_i to mean the $i-$th entry of a vector A. We frequently use i, j in the meaning of indices and with some abuse of notation we will use I, J to denote the index upper bounds. We further denote sets (and spaces) by Gothic letters ($\mathfrak{A}, \mathfrak{B}, \mathfrak{C}, \ldots$). Finally we often write $\mathbb{N}_I$ to denote the set $\{1, \ldots, I\}$.

2 Multilinear Estimation with Nuclear Norm Penalties

Recent research in statistics and machine learning [18] focused on composite norms. Regularization via composite norms allows one to convey specific structural a-priori information about the model to be estimated. Let $\mathcal{X} \in \mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2} \otimes \cdots \otimes \mathbb{R}^{I_N}$ denote a generic tensor. Consider the function:

$$g(\mathcal{X}) := \frac{1}{N} \sum_{n \in \mathbb{N}_N} \|\mathcal{X}_{<n>}\|_*$$

where $\cdot_{<n>}$ denotes the $n-$th unfolding operator and $\|\cdot\|_*$ is the nuclear norm for matrices. It can be shown that g is a well defined norm — that by extension can be called *nuclear* — and hence we write $\|\mathcal{X}\|_* := g(\mathcal{X})$. Furthermore such a norm represents an instance of a more general class that extends the concept of Shatten norms from matrices to higher order tensors [15]. Let $A : \mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2} \otimes \cdots \otimes \mathbb{R}^{I_N} \rightarrow \mathbb{R}^{D_1} \otimes \mathbb{R}^{D_2} \otimes \cdots \otimes \mathbb{R}^{D_M}$ be some linear map and assume $\mathcal{Z} \in \mathbb{R}^{D_1} \otimes \mathbb{R}^{D_2} \otimes \cdots \otimes \mathbb{R}^{D_M}$. In here we deal with the equality constrained optimization problem:

$$\hat{\mathcal{X}} := \underset{\mathcal{X} \in \mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2} \otimes \cdots \otimes \mathbb{R}^{I_N}}{\arg\min} f(\mathcal{X}) + \mu \|\mathcal{X}\|_* \tag{1}$$

$$\text{subject to} \quad A(\mathcal{X}) = \mathcal{Z} \tag{2}$$

aimed at finding a compact[1] $N-$order tensor-model $\hat{\mathcal{X}}$ based upon an application-dependent convex and smooth function f and a finite trade-off parameter $\mu > 0$. Algorithmically a solution of the unconstrained problem corresponding to (1) can be found by generating a sequence of convex and separable proximal problems [15] each of which can be solved via the Alternating Direction Method of Multipliers [5]. On the other hand a simple approach to deal with the linear constraint (2) is by means of a penalty method [2]. Interestingly the approach we propose is essentially a first order scheme in the tensor unknown. Hence its convergence speed can be improved relying on the concept of *estimating sequences* that underlies many recent proposals for l_1 and nuclear norm optimization [3],[16]. More details on the proposed strategy can be found in [15]. Here we only remark that the formulation in (1)-(2) can be used to tackle a broad class of tasks: different specifications of f give rise to different estimation problems both supervised and unsupervised. Examples follow.

2.1 Penalized Empirical Risk Minimization

Suppose we are given K input-output pairs $\left\{\left(y_k, \mathcal{Z}^{(k)}\right) \in \mathfrak{Y} \times \mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2} \otimes \cdots \otimes \mathbb{R}^{I_N}\right\}_{k \in \mathbb{N}_K}$ where $\mathfrak{Y}$ denotes the output set. Given a convex *loss function* $l : \mathfrak{Y} \times \mathbb{R} \rightarrow \mathbb{R}^+$ the unconstrained optimization problem associated to (1) can be used for supervised learning as soon as we take

$$f(\mathcal{X}) = \sum_{i \in \mathbb{N}_K} l\left(y_k, \left\langle \mathcal{Z}^{(k)}, \mathcal{X} \right\rangle\right) . \tag{3}$$

This corresponds to extending the *penalized empirical risk minimization* approach used in machine learning to the case where the generic input pattern is represented as a tensor $\mathcal{Z}$ and the prediction is performed via the linear function $\langle \mathcal{Z}, \mathcal{X} \rangle$. This is useful in a number of applications such as, for instance, classification of human action from surveillance videos or quality assessment of batches in chemometrics.

2.2 Tensor Denoising and Completion

Suppose we want to recover a low-rank tensor $\mathcal{X} \in \mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2} \otimes \cdots \otimes \mathbb{R}^{I_N}$ such that $A(\mathcal{X})$ is close or even coincide to an observed $\mathcal{Z} \in \mathbb{R}^{D_1} \otimes \mathbb{R}^{D_2} \otimes \cdots \otimes \mathbb{R}^{D_M}$. In the simplest situation

[1] Here *compact* means with small multilinear ranks, see [15].

$\mathcal{Z} \in \mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2} \otimes \cdots \otimes \mathbb{R}^{I_N}$ is a given noisy tensor observation and we are interested in recovering its latent version $\hat{\mathcal{X}}$, assumed to be compact. In this case we let

$$f(\mathcal{X}) = \|\mathcal{X} - \mathcal{Z}\|_\star^2$$

where $\|\cdot\|_\star$ denotes some smooth norm, such as the Frobenius norm. The constraint in (2) can be used to further impose strong prior information over $\mathcal{X}$ or a transformation thereof $A(\mathcal{X})$. A popular case (well-studied for second-order tensors) is found for the case where $(A(\mathcal{X}))_j = x_{i_1^j i_2^j \cdots i_N^j}$ and $\mathcal{Z} \in \mathbb{R}^J$ is a vector of measurements corresponding to a subset of entries with indices in

$$\mathfrak{O} = \left\{ (i_1^j, \ldots, i_N^j) \in \mathbb{N}_{I_1} \times \cdots \times \mathbb{N}_{I_N} : j \in \mathbb{N}_J \right\} .$$

In the limit case of tensor completion [9] we take $f = 0$. More details as well as concrete examples can be found in [15].

3 Beyond Linearity: Operatorial Representations

The core idea of kernel methods [13] consists of mapping input points represented as vectors (first order tensors) $\{Z^{(k)}\}_{k \in \mathbb{N}_K} \subset \mathbb{R}^p$ into a feature space of l_2 sequences (well behaved infinite[2] dimensional vectors) by means of a *feature map* $\phi : \mathbb{R}^p \to l_2$. Standard algorithms can then be applied to find a linear model of the type $\langle X, \phi_Z \rangle_{l_2}$ [1]. Computation in finite time is ensured thanks to finite dimensional representations [17]. Moreover, since the feature map is normally chosen to be nonlinear, a linear model $\langle X, \phi_Z \rangle_{l_2}$ in the feature space corresponds to a nonlinear function of Z in the original input space $\mathbb{R}^p$.

For tensors, our proposal to go beyond linearity corresponds to representing a tensor $\mathcal{Z}$ as a infinite dimensional operator $\Phi_{\mathcal{Z}}$ in the same spirit of the traditional kernel formalism where Z is represented by ϕ_Z. This requires the definition of an appropriate mapping approach as well as the existence of finite dimensional representations for $\mathcal{X}$ — which is now infinite dimensional — in the linear model $\langle \mathcal{X}, \Phi_{\mathcal{Z}} \rangle$. In the following we begin by characterizing the feature space of infinite dimensional $N-$th order tensors to which $\Phi_{\mathcal{Z}}$ and $\mathcal{X}$ belong. Successively, we present a possible operatorial representation. We conclude with remarks concerning finite representations and convexity.

3.1 Tensor Product of Hilbert Spaces

Assume Hilbert spaces (HSs) $(\mathfrak{H}_1, \langle \cdot, \cdot \rangle_{\mathfrak{H}_1})$, $(\mathfrak{H}_2, \langle \cdot, \cdot \rangle_{\mathfrak{H}_2}), \ldots,$ $(\mathfrak{H}_N, \langle \cdot, \cdot \rangle_{\mathfrak{H}_N})$. A space of infinite dimensional $N-$th order tensors can be constructed as follows. We recall that $\psi : \mathfrak{H}_1 \times \mathfrak{H}_2 \times \cdots \times \mathfrak{H}_N \to \mathbb{R}$ is a bounded (equivalently continuous) multilinear functional [8], if it is linear in each argument and there exists $c \in [0, \infty)$ such that $|\psi(h_1, h_2, \ldots, h_N)| \leq c \|h_1\|_{\mathfrak{H}_1} \|h_2\|_{\mathfrak{H}_2} \cdots \|h_2\|_{\mathfrak{H}_N}$ for all $h_i \in \mathfrak{H}_i$, $i \in \mathbb{N}_N$. It is said to be *Hilbert-Schmidt* if it further satisfies

$$\sum_{e_1 \in \mathfrak{E}_1} \sum_{e_2 \in \mathfrak{E}_2} \cdots \sum_{e_N \in \mathfrak{E}_N} |\psi(e_1, e_2, \ldots, e_N)|^2 < \infty$$

for one (equivalently each) orthonormal basis $\mathfrak{E}_i$ of $\mathfrak{H}_i$, $i \in \mathfrak{N}_N$. It can be shown that the collections of such well behaved Hilbert-Schmidt functionals endowed with the inner product

$$\langle \psi, \xi \rangle_{HSF} := \sum_{e_1 \in \mathfrak{E}_1} \sum_{e_2 \in \mathfrak{E}_2} \cdots \sum_{e_N \in \mathfrak{E}_N} \psi(e_1, e_2, \ldots, e_N) \xi(e_1, e_2, \ldots, e_N)$$

forms a HS. In particular, any multilinear functional associated to a $N-$tuple $(h_1, h_2, \ldots, h_N) \in \mathfrak{H}_1 \times \mathfrak{H}_2 \times \cdots \times \mathfrak{H}_N$ and defined by

$$\psi_{h_1, h_2, \ldots, h_N}(f_1, f_2, \ldots, f_N) := \langle h_1, f_1 \rangle_{\mathfrak{H}_1} \langle h_2, f_2 \rangle_{\mathfrak{H}_2} \cdots \langle h_N, f_N \rangle_{\mathfrak{H}_N} \tag{4}$$

belongs to such. Furthermore it can be shown that

$$\langle \psi_{h_1, h_2, \ldots, h_N}, \psi_{g_1, g_2, \ldots, g_N} \rangle_{HSF} = \langle h_1, g_1 \rangle_{\mathcal{H}_1} \langle h_2, g_2 \rangle_{\mathcal{H}_2} \cdots \langle h_N, g_N \rangle_{\mathcal{H}_N} . \tag{5}$$

[2] We are considering here the most general case associated, for instance, to the popular RBF Gaussian kernel.

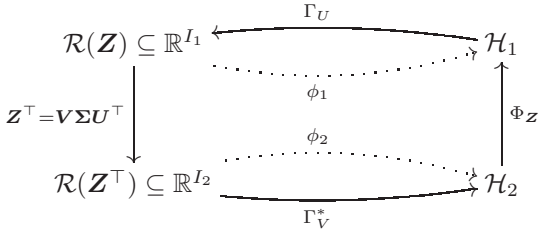


Figure 1: A diagram illustrating the operatorial representation for the second order case. The operator $\Phi_{\boldsymbol{Z}} \in \mathcal{H}_1 \otimes \mathcal{H}_2$ is the feature representation of the input pattern $\boldsymbol{Z} \in \mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2}$. With Γ_V^* we denoted the adjoint of Γ_V.

Starting from (4) we now let

$$h_1 \otimes h_2 \cdots \otimes h_N := \psi_{h_1,h_2,\ldots,h_N} \tag{6}$$

and define the tensor product space $\mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \cdots \otimes \mathcal{H}_N$ as the completion of the linear span

$$\operatorname{span}\{h_1 \otimes h_2 \otimes \cdots \otimes h_N \ : \ h_i \in \mathcal{H}_i,\ i \in \mathbb{N}_N\} \ .$$

A finite-rank element $\mathcal{X}$ of this space admit a representation in terms of a finite number J of rank-1 terms (6):

$$\mathcal{X} = \sum_{j \in \mathbb{N}_J} h_{i_1}^j \otimes h_{i_2}^j \cdots \otimes h_{i_N}^j \tag{7}$$

and can be envisioned as the infinite dimensional analogue of the traditional finite-rank tensors of previous section. If now $\mathcal{Y} = \sum_{j \in \mathbb{N}_R} g_{i_1}^r \otimes g_{i_2}^r \cdots \otimes g_{i_N}^r$, it follows from (5) that the inner product between $\mathcal{X}$ and $\mathcal{Y}$, denoted by $\langle \mathcal{X}, \mathcal{Y} \rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \cdots \otimes \mathcal{H}_N}$, is given by

$$\langle \mathcal{X}, \mathcal{Y} \rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \cdots \otimes \mathcal{H}_N} = \sum_{j \in \mathbb{N}_J} \sum_{r \in \mathbb{N}_R} \langle h_{i_1}^j, g_{i_1}^r \rangle_{\mathcal{H}_1} \langle h_{i_2}^j, g_{i_2}^r \rangle_{\mathcal{H}_2} \cdots \langle h_{i_N}^j, g_{i_N}^r \rangle_{\mathcal{H}_N} \ .$$

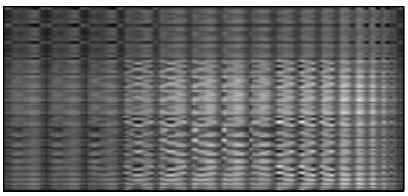
We further have that $\|\mathcal{X}\|_{\mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \cdots \otimes \mathcal{H}_N} = \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \cdots \otimes \mathcal{H}_N}}$.

Finally we stress that the present notion of tensor product should not be confounded with the one introduced in the context of *splines* [17],[4] and giving rise to functional ANOVA models [7]. In the latter case a tensor product formalism is used as a way of defining multivariate functions starting from univariate ones. Object in their tensor product space are then functions of the type $f : \mathbb{R}^d \to \mathbb{R}$ rather than operators, as in the present setting. A deeper look at the relation between the two constructions can be found in [12, Chapter 1.5].

3.2 Operatorial Representations



(a) A 19×18 grayscale image $\boldsymbol{Z}$ of a character taken from a natural scene.



(b) Its 190×171 feature representation $\Phi_{\boldsymbol{Z}}$.

Figure 2: An image $\boldsymbol{Z}$ (a) and its finite dimensional operatorial representation $\Phi_{\boldsymbol{Z}}$ (b) [14]. Here we used $2-$degree polynomial feature maps to generate the mode operators in (9).

Given the operatorial feature space sketched above it remains to define an appropriate feature representation $\Phi_{\mathcal{Z}}$ associated to a generic pattern $\mathcal{Z} \in \mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2} \otimes \cdots \otimes \mathbb{R}^{I_N}$. Here we follow [14] and restrict ourselves to the case of second order tensors. Hence we assume that we have input patterns represented as matrices $\{\boldsymbol{Z}^{(k)}\}_{k \in \mathbb{N}_K} \subset \mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2}$. The general case can be treated based upon the higher order analogues of the SVD [6]. Recall that the thin SVD decomposition of $\boldsymbol{Z} \in \mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2}$ can be written as

$$\boldsymbol{Z} = \sum_{i \in \mathbb{N}_r} \sigma_i U_i \otimes V_i \tag{8}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_{\min\{I_1,I_2\}} = 0$ are the ordered singular values and $U_i \otimes V_i$ are rank-1 matrices that represent the finite dimensional second-order analogue of (6). Let $\phi_1 : \mathbb{R}^{I_1} \to \mathcal{H}_1$ and $\phi_2 : \mathbb{R}^{I_2} \to \mathcal{H}_2$ be some feature maps in the standard sense of kernel methods. Based upon $\{U_i\}_{i\in\mathbb{N}_r}$ and $\{V_i\}_{i\in\mathbb{N}_r}$ we introduce the *mode*-0 operator $\Gamma_U : \mathcal{H}_1 \to \mathbb{R}^{I_1}$ and the *mode*-1 operator $\Gamma_V : \mathcal{H}_2 \to \mathbb{R}^{I_2}$ defined, respectively, by

$$\Gamma_U h = \sum_{i\in\mathbb{N}_r} \langle \phi_1(U_i), h\rangle_{\mathcal{H}_1} U_i \;\text{ and }\; \Gamma_V h = \sum_{i\in\mathbb{N}_r} \langle \phi_2(U_i), h\rangle_{\mathcal{H}_1} V_i \,. \tag{9}$$

Let $\Gamma_U \otimes \Gamma_V$ denotes the infinite dimensional analogue of the Kronecker product between matrices. We define the operatorial representation of $\boldsymbol{Z}$, denoted as $\Phi_{\boldsymbol{Z}}$, by

$$\Phi_{\boldsymbol{Z}} := \arg\min \left\{ \|\Psi_{\boldsymbol{Z}}\|^2_{\mathcal{H}_1\otimes\mathcal{H}_2} \;:\; (\Gamma_U \otimes \Gamma_V)\Psi_{\boldsymbol{Z}} = \boldsymbol{Z},\; \Psi_{\boldsymbol{Z}} \in \mathcal{H}_1 \otimes \mathcal{H}_2 \right\} . \tag{10}$$

This way $\boldsymbol{Z}$ is associated to the unique minimum norm solution of an operatorial equation. Details can be found in [14]. A diagram illustrating this idea is reported on Figure 1. On Figure 2 we show the (finite dimensional) feature representation obtained for the case where ϕ_1 and ϕ_2 are polynomial feature maps.

3.3 Conclusions: Finite Dimensional Kernel Representations and Practical Estimation

The generalized tensor-based framework that arise from the feature representation in (10) aims at combining the flexibility of kernel methods with the capability of exploiting structural information typical of tensor-based data analysis. The idea can be implemented into practical problem formulations [14] thanks to finite dimensional representations of the operatorial models. This is achieved via extensions of the classical Representer Theorem [17]. Unfortunately the current parametrization leads to non-convex optimization problems. Obtaining convex multilinear formulations within this framework is the subject of ongoing research.

Acknowledgments

Research supported by Research Council KUL: GOA Ambiorics, GOA MaNet, CoE EF/05/006 Optimization in Engineering(OPTEC), CIF1 and STRT1/08/023 IOF-SCORES4CHEM. Flemish Government: FWO: PhD/ postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0427.10N, G.0302.07 (SVM/Kernel), G.0588.09 (Brain-machine) research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940).

References

[1] M. Aizerman, E. M. Braverman, and L. I. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821 – 837, 1964.

[2] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear programming: theory and algorithms*. John Wiley and Sons, 2006.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[4] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

[5] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and distributed computation*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[6] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[7] C. Gu. *Smoothing spline ANOVA models.* Springer, 2002.

[8] R. V. Kadison and J. R. Ringrose. *Fundamentals of the theory of operator algebras*, volume 1. 1983.

[9] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. In *IEEE International Conference on Computer Vision (ICCV), Kyoto, Japan*, pages 8, 2009, 2009.

[10] I. Necoara and J. A. K. Suykens. Interior-Point Lagrangian Decomposition Method for Separable Convex Optimization. *Journal of Optimization Theory and Applications*, 143(3):567–588, 2009.

[11] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(\frac{1}{k^2})$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983, 1983.

[12] RA Ryan. *Introduction to tensor Product of Banach Spaces*. Springer-Verlag New York, LLC, 2002.

[13] B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[14] M. Signoretto, L. De Lathauwer, and J. A. K. Suykens. Kernel-based learning from infinite dimensional 2-way tensors. In *ICANN 2010, Part II, LNCS 6353*, 2010.

[15] M. Signoretto, L. De Lathauwer, and J. A. K. Suykens. Nuclear Norms for Tensors and Their Use for Convex Multilinear Estimation. *Internal Report 10-186, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), Lirias number: 270741*, 2010.

[16] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2008.

[17] G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1990.

[18] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37:3468–3497, 2009.