Matroid Regression

Franz J. Király * Louis Theran†

Abstract

We propose an algebraic combinatorial method for solving large sparse linear systems of equations locally - that is, a method which can compute single evaluations of the signal without computing the whole signal. The method scales only in the sparsity of the system and not in its size, and allows to provide error estimates for any solution method. At the heart of our approach is the so-called regression matroid, a combinatorial object associated to sparsity patterns, which allows to replace inversion of the large matrix with the inversion of a kernel matrix that is constant size. We show that our method provides the best linear unbiased estimator (BLUE) for this setting and the minimum variance unbiased estimator (MVUE) under Gaussian noise assumptions, and furthermore we show that the size of the kernel matrix which is to be inverted can be traded off with accuracy.

1. Introduction

Sparse linear systems are a recurring topic in modern science. They occur in a wide variety of contexts such as numerical analysis, medical imaging, control theory or signal processing. Of particular interest are sparse linear systems of big size - that is, a number of equations which is of the order of thousands, millions, billions or more - since they occur in practice, e.g. in linear inverse problems such as tomography, or analysis of large scale data with sparse structure as they occur in recommender systems or network analysis.

Whole areas of research, spanning disciplines in most areas of science, have been devoted to the end of solving linear systems of equations $Ax = b$ where A is huge and sparse. A selection of books on the topic, in which numerical solution strategies are outlined, and which is far from being representative, includes [1, 4, 5, 10, 11]. Further important is the area in medical imaging which is concerned with sparse linear systems specifically arising from certain geometries in tomography, compare the algebraic approaches in [6, 7], for which specific techniques have been developed. Moreover, we would like to mention that sparse matrices and their spectral properties also appear as a recurring topic in networks, see e.g. [2].

Regarding the huge corpus of existing literature, we would, however, like to stress one fact: the state-of-the-art methods and theories mostly make use of spectral or analytical properties of the huge matrices; efficient methods which use particular structure - be it algebraic or combinatorial - of the sparse system of equations, seem not to be available. Furthermore, all methods

*Department of Statistical Science, Univerity College London, and MFO f.kiraly@ucl.ac.uk
†Inst. Math., AG Diskrete Geometrie, Freie Universität Berlin, theran@math.fu-berlin.de

usually seek a complete solution of the system in terms of x, while scenarios where some projection Px with P a matrix and Px of tractable size might be desirable - e.g., if in the tomography scenario, only part of the scanned region needs a high resolution, or in the networks scenario, where only part of the network might be of, say, predictive interest. Similarly, in the recommender systems scenario, it is more natural to make a recommendation for a single item instead of making all possible recommendations at once.

In this paper, we propose theoretical foundations and practical methods to address this kind of problem, which have the potential advantage of scaling with the row-size of P instead of the size of A. That is, optimally the method will have a running time that does not scale with the size of A, only with certain sparsity properties of A which in many practical scenarios scale constant with respect to the size of A. The only assumption we will need for this to work is that there exist a sufficient number of linear dependencies of rows of A which are sparse in their coefficient representation. This is frequently the case if A has intrinsic combinatorial meaning, or is highly structured otherwise.

The central ingredient is the notion of regression matroid, which provides a kind of dictionary for minimal such dependencies (= circuits), and the circuit kernel matrix, which is the covariance matrix between the circuits. Restricting to small circuits in a "neighborhood" of P, we are able to obtain a least squares estimator for Px where the most costly ingredient is inversion of the circuit kernel matrix - which scales with the size and number the circuits, and not the size of A. Therefore, through choosing the circuits - interpreted as the “locality” parameter of matroid regression - we also obtain a tool of trading off accuracy of the solution with computational cost.

More concisely, our main contributions are:

- the notions of **matroid regression** and **circuit kernel**, capturing algebraic combinatorial properties of the linear system
- an **explicit algorithm** computing a variance minimizing estimator for the evaluation Px
- an explicit form for the **variance of that estimator** which depends not on x but only on the noise model
- an **explicit algorithm** to compute that variance without computing Px
- **proofs of optimality** and universality for the estimator (BLUE in general, MVUE for Gaussian noise and for unknown noise)
- a proof of the error being monotonous in the “locality” of the estimate, yielding a **complexity-accuracy-tradeoff**
- **characterization of the regression matroid** in some cases, including potential measurements, 2-sparse vectors, rank one matrix completion; explanation how in these cases circuits and combinatorial properties of **characteristic graphs** relate

Our framework also explains some particular findings in the case of matrix completion [8], which we can reproduce by reduction to a sparse linear system, and solves open questions about the optimality of the estimators raised in [8]. In the same sense, we hypothesize that the matroid regression methods have a rather general and natural extension to the non-linear case.

2. Structured linear estimation

We will consider two compressed sensing problems which are dual to each other. In the sequel, the field $\mathbb{K}$ is always one of $\mathbb{C}$ or $\mathbb{R}$, and the parameter n will be the signal dimension.

Problem 2.1 (Primal problem **(P)**)**.** There is an unknown signal $x \in \mathbb{K}^n$ observed via a *linear measurement process*:

$$b = Ax + \varepsilon \tag{1}$$

The noise ε is centered and has finite variance, and the matrix $A \in \mathbb{K}^{N\times n}$ is known. The task is to compute a *linear evaluation* $\gamma = \langle w, x\rangle$, for a known $w \neq 0$ in the row-span span$A$.

In general, n will be large, and potentially $N \gg n$, but A will be either *sparse*, *structured* or both. This means that simply inverting an $n \times n$ sub-matrix of A is not a good solution. Instead, we will show how to use the structure of A to find solutions *locally*, using very few coordinates of b or both.

Problem 2.2 (Dual problem **(D)**)**.** There is an unknown signal $y \in \mathbb{K}^n$ and an unknown scalar γ, satisfying a constraint

$$\gamma w' = A'y \tag{2}$$

The task is to estimate γ from observations $b = y + \varepsilon$, with $A' \in \mathbb{K}^{N\times n}$ and w' known.

Since the Dual Problem (D) can be treated with the same methods, we will focus on the Primal problem (P).

2.1. The Problems in Context

We interpret the general problem (P) as a *supervised learning* problem. To see this, take the rows $a_1, a_2, \ldots, a_N$ to be training data points and the coordinates $b_1, \ldots, b_N$ of b to be training labels. The unknown vector x is then the regressor, and the learning tasks can be: (i) imputation of single coordinates of x; (ii) prediction of the label of a new point w; (iii) denoising, which corresponds to w being one of the a_i; among others.

Alternatively, even though N is typically quite large, so that x is not compressed in the classic sense, problem (P) can be interpreted in terms of *compressed sensing*. Here, the task is to use as few coordinates of b as possible to estimate $\langle w, x\rangle$ accurately. This should be contrasted with the approach of computing the (pseudo-)inverse of A, e.g., for i.i.d. noise the estimator $w^\top A^{-1} b$.

2.2. Example Instances

To fix, the concept, we show how to cast some scenarios in terms of problem (P).

Example 2.3 (Measuring potentials)**.** The task is to measure from an unknown potential x, given a set of measurements. The rows of A are of the form $e_j - e_i$, where $\{e_i : i \in [n]\}$ are the standard basis vectors of $\mathbb{K}^n$. The vector w is also of this form.

Example 2.4 (Rank 1 Matrix Completion)**.** The task is to impute or denoise the entry at position (i, j) in a partially-observed, $m \times n$ rank 1 matrix $\mathbf{A} = \mathbf{u} \cdot \mathbf{v}^\top$. The vector x is the concatenation of the entry-wise logarithms of $\mathbf{u}$ and $\mathbf{v}$; the vector $e_{j+m} - e_i$ is a row of A if the position (i, j) is observed; the vector b is the vectorization of the set of observed entries; $w = e_j - e_i$ where (i, j) is the position of the entry to impute or de-noise.

Example 2.5 (Discrete tomography)**.** The task is to reconstruct a bitmap image (subset of a lattice in Euclidean space) from a number of projections. The matrix A has a decaying spectrum and w describes a regularized region of interest.

3. Regression matroids

Our strategy for solving the problem (P) will be to exploit the *structure* of the constraint matrix A. The object that captures this is the *regression matroid* of A and w, which we now define.

Definition 3.1. Let $a_1, \ldots, a_N \in \mathbb{K}^n$ be a collection of vectors, let $w \in \mathbb{K}^n$ be a target vector.

(i) The (linear) *regression matroid* associated to the a_i and w is the pair $([N], \mathcal{I})$, where

$$\mathcal{I} := \big\{ I \subseteq [N] : \text{the set } \{w\} \cup \{a_i : i \in I\} \text{ is linearly independent} \big\}$$

and we write $a_* := w$. We will denote the matroid by $L(w|a_1, \ldots, a_N) := ([N], \mathcal{I})$. If A is the matrix having a_i as i-th row, we simply write $L(w|A)$.

(ii) A set $C \subseteq [N]$ with $* \in C$ is called (linear) *particular regression circuit* of $L(w|A)$, if the equation $w = \sum_{i \in C} \lambda_i a_i$ implies $\lambda_i \neq 0$ for all $i \in C$.

(iii) A set $C \subseteq [N]$ is called *general regression circuit* of $L(w|A)$, if it is a regression circuit of $L(0|A)$.

In matroid terms, the regression matroid is the elementary quotient of the linear matroid of A by the element w. Note that a set $C \subseteq [N]$ can not be both a particular and general regression circuit. Also, if w is one of the a_i, then $\{i\}$ is a particular regression circuit. An extension to more than one target vector is straightforward, but for simplicity, we continue with only the single target vector w.

If $\lambda \in \mathbb{K}^n$ is a vector, we say that the *support* of λ is the set $\{i \in [N] : \lambda_i \neq 0\}$. Circuits and regression circuits correspond to linear dependencies with minimal support.

Proposition 3.2. *Let $L(w|A)$ be a regression matroid. Then:*

(i) $C \subseteq [N]$ is a particular regression circuit if and only if there is a unique vector $\lambda \in \mathbb{K}^N$ supported on C such that $w = \lambda A$.

(ii) $C \subseteq [N]$ is a general regression circuit if and only if there is a unique, up to scalar multiplication, vector $\lambda \in \mathbb{K}^n$ supported on C such that $\lambda A = 0$.

Proof. We will prove (i), since the proof of (ii) is similar. Suppose that $\lambda_1 A = w$ and $\lambda_2 A = w$ on a set C, and let $j \in [N]$ be arbitrary. Set $\alpha = \lambda_1(j)/\lambda_2(j)$. Then $(\alpha\lambda_2 - \lambda_1)A = (\alpha - 1)w$, and the support of $(\alpha\lambda_2 - \lambda_1) \subsetneq C$. In particular, C has minimal support if and only if $\alpha = 1$. Since j was arbitrary, we are done. □

Proposition 3.2 justifies the following definition:

Definition 3.3. Let $w \in \mathbb{K}^n$, let $A \in \mathbb{K}^{N \times n}$, let C be a particular regression circuit of $L(w|A)$. We call the unique vector λ associated to C by Proposition 3.2 the *circuit vector* of C. Similarly, we may pick a normed representative to define the *circuit vector* of a general circuit. For a circuit C of either type (recall that a circuit can be particular or general, but not both), we will use λ_C to denote its circuit vector.

We introduce a definition for formal linear combinations of circuits:

Definition 3.4. For circuits $C_1, \ldots, C_m$ and $\alpha_1, \ldots, \alpha_m \in \mathbb{K}$, we will define a *circuit divisor* to be a formal linear combination

$$\alpha C_1 + \cdots + \alpha_m C_m,$$

and associate to it the circuit vector $\alpha_1 \lambda_{C_1} + \cdots + \alpha_m \lambda_{C_m}$. We denote the $\mathbb{K}$-vector space of all circuit divisors of $L(w|A)$ by $\mathcal{C}(w|A)$, and we write $0 = 0C$. Two circuit divisors D_1, D_2 are called *linearly equivalent* if their circuit vectors are the same, in which case we write $D_1 \sim D_2$.

The purpose of this notation is to put an emphasis on the algorithmic process of combining circuits, over the pure consideration of the circuit vector. Indeed, in general, the same circuit vector can be obtained from different formal linear combinations of circuits.

The principal objects for solving the sparse linear system are the respective spans of particular and general regression circuits, which we will term regression space and general circuit space. They can be seen as analogues to the particular and general solutions occurring in the theory of differential equations: for solving the linear system accurately and efficiently, we need to find one particular regression circuit in the regression space, and a sufficient number of general circuits in the general circuit space.

Definition 3.5. Let $L(w|A)$ be a regression matroid. The *regression space*, is affine span

$$\mathcal{C}_p(w|A) := \text{aff}\{C \in \mathcal{C}(w|A) : C \text{ is a particular regression circuit}\}$$

(where aff denotes the affine hull), and the *general circuit space* is

$$\mathcal{C}_c(w|A) := \text{span}\{C \in \mathcal{C}(w|A) : C \text{ is a general regression circuit}\}$$

Elements of $\mathcal{C}_p(w|A)$ are called *particular regression divisors*, elements of $\mathcal{C}_c(w|A)$ are called general regression divisors.

Since $\mathcal{C}_p(w|A)$ contains circuit divisors, and not just the circuit vectors, is has richer structure than the left kernel A.

The relationship between the two spaces is:

Lemma 3.6. *Let $L(w|A)$ be a regression matroid with $w \neq 0$. Then:*

(i) $\text{Ker}A = \{\lambda_D \ : D \in \mathcal{C}_c(w|A)\}$

(ii) *Let C_p be a fixed particular regression circuit vector. Then*
$\mathcal{C}_p(w|A) = \{C : C \sim C_p + D \text{ with } D \in \mathcal{C}_c(w|A)\}$

Proof. The first equality follows from the fact that the kernel vectors with minimal support span the kernel. For the second, it suffices to prove that for two particular regression circuits C_1, C_2, there is $D \in \mathcal{C}_c(w|A)$ such that $C_1 - C_2 \sim D$. By definition, $(\lambda_{C_1} - \lambda_{C_2})A = 0$, therefore $\lambda_{C_1 - C_2} \in \text{Ker}A$ the statement then follows from the first equality. □

One could interject that representing the left kernel vectors of A in terms of circuits and circuit divisors is unnecessarily complicated. The theoretical estimator will be formulated, in terms of the non-zero entries of the circuit vector λ_C; also, the algorithmic procedure will also benefit from treating the circuits as sets of indices instead of the circuit vector λ_C. The reader is invited to think about circuits and divisors simultaneously in terms of the circuit vectors plus the

information which entries are non-zero, but we think that the notion of circuit divisors makes more clear where the advantages of our algebraic combinatorial method lie.

The final object we need to define before describing the estimation procedure is linear spans of divisors:

Definition 3.7. (i) A $\mathbb{K}$-vector space $\mathcal{L}_c \subseteq \mathcal{C}_c(w|A)$ closed under linear equivalence is called *linear system of general circuit divisors,* or short, *general system* of circuits.

(ii) A $\mathbb{K}$-affine space $\mathcal{L}_p \subseteq \mathcal{C}_p(w|A)$ closed under linear equivalence is called *affine system of particular circuit divisors*, or short, *particular system* of circuits.

A general/particular system of circuit divisors $\mathcal{L}$ is said to be generated by circuit divisors $C_1, \ldots, C_m$ if every element in $\mathcal{L}$ is linearly equivalent to a linear/affine combination of the C_i. In this case, the $C_1, \ldots, C_m$ are called *generating system* of $\mathcal{L}$, and if m is additionally minimal, they are called a *basis* of $\mathcal{L}$ (in both the linear/affine cases).

Remark 3.8 — An important example of particular systems is given as follows: let $\mathcal{L}_c$ be a general system, and $C_p \in \mathcal{C}_p(w|A)$, for example C_p a particular regression circuit. Then, $C_p + \mathcal{L}_c := \{C_p + C_c \,:\, C_c \in \mathcal{L}_c\}$ is a particular system.

Particular systems will be one of the main ingredients in estimating the projection $\langle w, x\rangle$. The above remark shows that to this end, it suffices to acquire a single particular circuit and some general system of circuits.

Examples of regression matroids We give some examples of regression matroids and discuss the structure of their regression circuits.

Example 3.9 (Uniform regression matroid)**.** If A is generic, then $L(w|A)$ is a quotient of a rank n uniform matroid, so all special regression

Example 3.10 (Generic low-rank)**.** If A is generic of rank r, any r rows will form a special regression circuit, any $r+1$ rows a general one. In particular, if $r \ll n$, then special regression circuits have *sparse* support. only rank r. As in Example 3.9, the special regression circuits are easy to find, but now they are *sparse*, provided $r \ll n$: only r rows are required.

Example 3.11 (Graphic regression matroids)**.** as a basis for the space of cycles. The regression space is therefore equivalent to the first homology of the graph G, a basis of which can be efficiently computed in $O(n+N)$ time.

The potentials Example 2.3 and matrix completion Example 2.4 both give rise to graphic regression matroids. We will explore strategies for finding good sets of special regression circuits in this case below.

Another combinatorial example comes from matrices with *sparse filling patterns* and *generic* non-zero entries.

Example 3.12 ($(1,\ell)$-sparsity matroids)**.** We now define a kernel which will be key in our estimation procedure. The intuition behind the technical definition is that special regression circuit vectors define linear If the rows of A have at most d generic non-zero entries, and the kernel of A is spanned by $\ell \leq d-1$ generic vectors, then $L(w|A)$ is the quotient of a $(1,\ell)$-sparsity matroid on a d-hypergraph.

4. Matroid Regression

The strategy for constructing the matroid regression estimator is as follows: each special regression divisor produces one exact estimate for the evaluation $\langle w, x\rangle$. The estimator is obtained for the choice of regression divisor minimizing variance. Since the special regression divisors form an affine space, on which variance is a quadratic form, we obtain the variance minimizing estimate as explicit solution to a quadratic system. The major algorithmical advantage of the matroid regression view was outlined in Remark 3.8: after finding one special regression circuit, general regression circuits that are easier to find can be used to produce more special circuits in order to decrease the variance and this the estimation error.

4.1. An Unbiased Estimator

Recall problem (P): we are provided with the data for a regression matroid $L(w|A)$, with $A \in \mathbb{K}^{N\times n}$ known and $x \in \mathbb{K}^n$ unknown, and want to estimate an evaluation $\langle w, x\rangle$ of the unknown signal x, from $b = Ax + \varepsilon$. Let Σ be the covariance matrix of the N-dimensional random vector ε. First we construct circuit-vector estimators.

Proposition 4.1. *Let D be a particular regression divisor of $L(w|A)$, with circuit vector λ_D. Then,*

$$\widehat{\gamma}(D) := \langle \lambda_D, b\rangle$$

is an unbiased estimator for $\langle w, x\rangle$ with variance $\mathrm{Var}(\widehat{\gamma}(D)) = \lambda_D^* \Sigma \lambda_D$. Conversely, all unbiased estimators linear in b are of the type $\widehat{\gamma}(D)$ for some particular regression divisor D.

Proof. By linearity of expectation and centeredness of ε, it follows that

$$\mathbb{E}(\widehat{\gamma}(D)) = \langle \lambda_D, \mathbb{E}(b)\rangle = \lambda_D A x = \langle w, x\rangle,$$

where the last equality follows from the fact that λ_D is a circuit divisor - thus $\widehat{\gamma}(D)$ is unbiased. The statement for the variance follows from bilinearity of covariance via the equation

$$\mathrm{Var}(\widehat{\gamma}(D)) = \mathrm{Var}(\langle \lambda_D, b\rangle) = \lambda_D^* \mathrm{Var}(b)\lambda_D = \lambda_D^* \mathrm{Var}(\varepsilon)\lambda_D.$$

The converse statement follows from Lemma 3.6. □

Proposition 4.1 shows how a good estimator for $\langle w, x\rangle$ can be obtained: find a divisor D with small variance. Since the latter is quadratic in λ_D, this can be reduced to a quadratic optimization problem. However, there is one major issue with the present formulation: such an optimization would be essentially over λ_D, not in terms of the circuits, and eventually involve inversion of an $(N \times N)$ matrix - therefore nothing is gained yet with respect to the pseudo-inversion done in usual linear regression. To address this issue, we will express the variance from Proposition 4.1 in terms of circuits and divisors.

4.2. The Circuit Kernel

The circuit kernel is the analogue of the covariance matrix of the estimator $\widehat{\gamma}$, but represented in the coordinates induced by circuits and formal divisors. It yields a quadratic form on the circuit space $\mathcal{C}(w|A)$, allowing optimization to take place over the combinatrial structure of the circuits as compared to the circuit vectors λ_*.

Definition 4.2. Fix a covariance matrix $\Sigma \in \mathbb{K}^{N\times N}$. For two regression divisors D_1, D_2 with circuit vectors λ_1, λ_2, we define the *circuit kernel function*

$$k(D_1, D_2) = \lambda_1^* \Sigma \lambda_2.$$

For a collection $D_1, \dots, D_m$ of regression divisors and Σ, we define the *circuit kernel matrix* K to be the $(m \times m)$ matrix which has $k(D_i, D_j)$ as entries.

Lemma 4.3. *The circuit kernel is a positive semi-definite bilinear form on $\mathcal{C}(w|A)$. For a particular regression divisor $D \in \mathcal{C}_P(w|A)$, it holds that* $\mathrm{Var}(\widehat{\gamma}(D)) = k(D, D)$.

Proof. The matrix Σ is positive semi-definite as covariance matrix of a random variable. Therefore, there is a Cholesky decomposition $\Sigma = U^\top U$ with $U \in \mathbb{R}^{N\times N}$. Observe that by definition, any circuit kernel matrix K will be of the form $K = \Lambda^* \Sigma \Lambda$ for $\Lambda \in \mathbb{K}^{N\times m}$ and some m. Therefore, $K = (U\Lambda)^*(U\Lambda)$ is a positive semi-definite matrix, which implies positive semi-definiteness of k. The second statement follows from Proposition 4.1 and the definition of k. □

Proposition 4.4. *Let $\mathcal{L}$ be a particular system, generated by $C_1, \dots, C_m$. The quadratic form $k(D, D)$ is minimized for $D \in \mathcal{L}$ by exactly the divisors*

$$D = \sum_{i=1}^{n} \alpha_i C_i, \text{ where } \alpha \in \left(K^{-1}\mathbf{1}\right)\left(\mathbf{1}^\top K^{-1}\mathbf{1}\right)^{-1},$$

$\mathbf{1}$ *is the vector of ones, K is the $(m \times m)$ kernel matrix with entries $k(C_i, C_j)$, and $K^{-1}\mathbf{1} = \{x \in \mathbb{K}^n : Kx = \mathbf{1}\}$.*

Proof. Since $\mathcal{L}$ is a particular system, it holds that $D \in \mathcal{L}$ if and only if $\mathbf{1}^\top \alpha = 1$. Bilinearity of K implies that $k(D, D) = \alpha^\top K \alpha$. From this, we obtain the Lagrangian

$$L(\alpha, \xi) = \alpha^\top K \alpha + \xi\left(1 - \mathbf{1}^\top \alpha\right),$$

where the slack term models the condition $\mathbf{1}^\top \alpha = 1$. A straightforward computation yields

$$\frac{\partial L}{\partial \alpha} = 2K\alpha - \xi\mathbf{1}$$

By Lemma 4.3 K is positive semi-definite, therefore $\alpha^\top K \alpha$ is convex, so the minimizers of $k(D, D)$ satisfying $\mathbf{1}^\top \alpha = 1$ will exactly correpond to the $\alpha \in K^{-1}\mathbf{1}/\mathbf{1}^\top K\mathbf{1}$. □

4.3. The Optimal Estimator

We are now ready to give the final form of our estimator:

Theorem 4.1. *Let $\mathcal{L}$ be a particular system. Let $D \in \mathcal{L}$ be any divisor minimizing $k(D, D)$, as in Proposition 4.4. Consider the estimator*

$$\widehat{\gamma}(\mathcal{L}) := \langle \lambda_D, b \rangle.$$

(i) *$\widehat{\gamma}(\mathcal{L})$ is independent of the choice of the minimizer D of $k(D, D)$.*

(ii) $\widehat{\gamma}(\mathcal{L})$ *is an unbiased estimator for* $\langle w, x\rangle$.

(iii) $\operatorname{Var}\widehat{\gamma}(\mathcal{L}) = k(D,D) = \min_{D\in\mathcal{L}} \operatorname{Var}\left(\widehat{\gamma}(D)\right)$.

Proof. (i) follows from elementary linear algebra. (ii) follows immediately from Proposition 4.1. (iii) follows from Lemma 4.3. □

Theorem 4.1 indicates an algorithmic way to obtain good estimates for $\langle w, x\rangle$: namely, first find generators $C_1, \ldots, C_m$ for a particular system; then determine the minimizer D as described in Proposition 4.4, keeping track of the circuit vector. Finally, compute $\widehat{\gamma}$. At the same time, Theorem 4.1 highlights several important advantages of our estimator $\mathcal{L}$. First, computation of $\widehat{\gamma}$ involves only (pseudo-)inversion of an $(m \times m)$-matrix, as opposed to (pseudo-)inversion of an $(n \times n)$-matrix for the naive strategy - this is an advantage in the sparse setting, as we will show that m can be chosen small in some common scenarios. Second, Theorem 4.1 (i) in particular shows that the estimate does not depend on the particular generating system chosen for $\mathcal{L}$; therefore, following Remark 3.8, we may choose a system of the form $D_i = C_p + C_i$, where C_i are general circuits, and C_p is the same particular circuit for all D_i. This means, for each new w, we only need to find a single particular circuit, while the system of general circuits given by the C_i needs only to be changed when A changes. Due to the bilinear equality $k(D_i, D_j) = k(C_p, C_p) + k(C_p, C_i) + k(C_p, C_j) + k(C_i, C_j)$ this also means that the kernel matrix K has to be computed only once per A.

4.4. Algorithms

We provide algorithms computing the estimated evaluation and variance bounds for the error.

Since the the circuit vectors, the circuit kernel matrix K, and the optimal α are required for both, we first compute those, given a collection of circuits $C_1, \ldots, C_m$. Algorithm 1 outlines informal steps for this. We use MATLAB notation for submatrices and concatenation. The com-

Algorithm 1 Computes circuit kernel K and α.
Input: A, w, circuits $C_1, \ldots, C_m$, covariance matrix Σ.
Output: circuit union C, circuit vector matrix Λ, kernel matrix K and minimizer α.

1: For all $i = 1 \ldots m$, compute the circuit vector λ_i of the C_i as the normalized left kernel vector of the matrix $[A[C_i, :]; w]$
2: Write the λ_i as rows of a matrix Λ, with rows indexed by C.
3: Write $C = C_1 \cup \cdots \cup C_m$.
3: Compute the kernel matrix $K = \Lambda^* \cdot \Sigma[C, C] \cdot \Lambda$
4: Calculate $\alpha = \left(K^{-1}\mathbf{1}\right)\left(\mathbf{1}^\top K^{-1}\mathbf{1}\right)^{-1}$.
5: Output Λ, K and α.

putations in steps 1 and 3 can be done fairly efficiently, since while Σ or A may be huge, the circuits C_i select only small submatrices. In an optimal scenario, increasing the number of rows of A has only a small or negligible effect on the size of C. Note that inputting b is not required, therefore Algorithm 1 needs not to be rerun if A, w, Σ stay the same, i.e., if it is only the signal x which changes.

Algorithms 2 and 3 takes the computed invariants from Algorithm 1 and computes estimates for $\langle w, x\rangle$ and its variance. These algorithms consist only of multiplications, and again can be

made efficient by the fact that the occurring matrices are of size at most $\#C$, therefore again controlled by the choice of circuits.

Algorithm 2 Estimates the evaluation $\langle w, x\rangle$.
Input: w,A, b, a collection of circuits $C_1, \ldots, C_m$, covariances Σ.
Output: The variance-minimizing estimate $\widehat{\gamma}(\alpha)$ for $\langle w, x\rangle$.

1: Compute Λ and α with Algorithm 1.
2: Write $C = C_1 \cup \cdots \cup C_m$ (this could also be obtained from Algorithm 1)
3: Return $\alpha^\top \cdot \Lambda \cdot b[C]$ as an estimate.

Algorithm 3 Estimates the variance of the evaluation $\langle w, x\rangle$.
Input: A, w, a collection of circuits $C_1, \ldots, C_m$, covariances Σ.
Output: The variance lower bound for $\log(A_{ij})$.

1: Calculate K and α with Algorithm 1.
2: Return $\alpha^\top \cdot K \cdot \alpha$.

Algorithm 3 can be used to obtain the variance bound independently of the observations in b - therefore an error estimate which is independent of the algorithm which does the actual estimation.

We would further like to note that the size of all matrices multiplied or inverted in the course of all three algorithms is bounded by the cardinality of the circuit union C. The only matrix of potentially larger size is $[A[C_i, :]; w]$ which can have more columns than $\#C$, up to n. However, there is no noise on A, and this matrix is used only to compute the unique (up to multiplicative constant) left kernel vector, so it can be replaced by the matrix consisting of any $\#C + 1$ linearly independent columns. Therefore, once a suitable circuit basis $C_1, \ldots, C_m$ is known which is accurate enough, $\langle w, x\rangle$ can be estimated in complexity depending only on $\#C$, and not on N or n.

4.5. On Finding Circuits

While the algorithms presented in section 4.4 are fairly fast and near-optimal by the considerations in sections 5.2 and 5.3, they highly rely on the collection of circuits which is input and which determines the submatrix to consider. Therefore, one is tempted to believe that the difficult problem of inverting A has merely been reduced to a combinatorial problem which is more difficult. The point here is again that if A and w are sparse, or if there is different combinatorial structure implying small circuits, this combinatorial problem has a comparably simple solution in practical settings. For example, if not much is known about A, but it has small circuits that are well-dispersed, one can attempt to find circuits via ℓ^1-minimization, e.g., by solving the convex program

$$\min \|\lambda\| \quad \text{subject to} \quad \lambda^\top \cdot A[D] = 0,$$

where D is is a randomly chosen subset of a number of columns which is likely to contain a circuit. On the other hand, if the rows of A and/or w have a specific combinatorial structure, for example related to properties of graphs, this can open up the problem to efficient algorithms which scale

with the problem's sparsity instead of its size. One can regard the matrix completion algorithm from [8] as a proof of concept for this, since the computation of the graph homology may be done in a local neighborhood around the missing entry whose size is constant, we will explain this in more detail in section 4.6. We will also list more examples with different combinatorial features that can be treated in this way.

4.6. Example Cases

The algorithms outlined in section 2 provide a fast and stable way of computing the evaluation once enough circuits have been identified. One main advantage of our strategy is that for each matrix A, the circuits need to be computed only once, and can be applied for different signals x. Furthermore, if the sparse matrix A (or its dual A') is highly structured - as it frequently occurs when analyzing network structure - then so are the circuits, in which case they can be obtained by combinatorial algebraic methods. We list some basic examples for demonstration purposes.

The Sample Mean and Linear Regression Both sample mean and ordinary least squares regression can be recovered as special cases of matroid regression. The sample mean is obtained for setting A to be an N-vector of ones, $w = 1$ and Σ the identity matrix - regression circuits consist of exactly one element, with the circuit vector being the corresponding standard basis vector. Least squares regression is obtained for setting Σ to be the identity and estimating evaluations for $w = e_i, 1 \leq i \leq n$ with e_i being an orthonormal system for $\mathbb{K}^n$.

Multiple Observations A behavior related to sample mean can be observed if multiple copies of the same row occur in A. In this case, a regression circuit will contain exactly one of those, and there will be a special regression circuit of the same type for each of the copies. Furthermore, for each pair of copies, a general circuit will appear containing exactly that pair. In order to prevent multiplicative growth of the number of circuits, it is suggested to pool multiple observations in a single one by taking the covariance-weighted mean.

Denoising A related case is if A contains w as a row. Here, that row will occur as a special regression circuit with only one element. Applying matroid regression in this case will trade off the noise in that single observation through the relations with other rows, therefore can be interpreted as a denoising of that observation. Even w occurs multiple times as a row of A, matroid regression will in general improve over merely taking the covariance-weighted sample mean of those rows' observations.

Measuring Potentials We consider the case where x corresponds to a potential, and differences are measured. In this case the rows of A take the form $e_i - e_j$ with e_i the standard basis for $\mathbb{K}^n$; assume that $w = e_k - e_\ell$ is of the same form. Let G be the oriented graph with n nodes which has an edge (i, j) if and only if A has a row $e_i - e_j$. Then the following characterization for regression circuits and general circuits can be shown: a set of edges is a special regression circuit if and only if it forms a path from k to ℓ contained in G - including possibly the edge (k, ℓ) itself in case w occurs as a row of A. The corresponding circuit vector consists of ones. A set of edges is a general circuit if and only if it is a cycle contained in G. Small circuits can therefore be efficiently found by finding elements in the first graph homology of G around the edge (k, ℓ).

Sparse Sums The case where the rows of A are of form $e_i + e_j$, and $w = e_k + e_\ell$ is very similar. Let G be the (simple) graph with n vertices and the same edge assignment as above. In this case, the special regression circuits will be exactly paths of odd length from k to ℓ contained in G, with circuit vectors being alternatingly -1 and 1, starting with -1. General circuits will be cycles of even length, with circuit vectors alternatingly 1 and -1. As in the potentials case, a search of the first graph homology will provide cycles near (k, ℓ) efficiently.

Low-Rank Matrix Completion By taking logarithms, compare the general strategy in [8], the rank one matrix completion problem can be transformed to the following linear problem: write the true rank 1 matrix $X \in \mathbb{R}^{m \times n}$ ($X =$ the A from the cited paper) as $X = uv^\top$ with $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$. Then, x is an $(m+n)$-vector that is concatenation of component-wise $\log u$ and $\log v$. The rows of the matrix A consist of concatenations (e_i, e'_j) of standard basis vectors $e_i \in \mathbb{R}^m$ and $e'_j \in \mathbb{R}^n$, being present if the entry (i, j) is observed; w is of the same form, corresponding to the unobserved entry (k, ℓ). This exposes rank one matrix completion as a sub-case of the "sparse sums" scenario discussed above. Note that the graph G is always bipartite due to how A was constructed, and that the missing entry of the matrix can be completed from a local neighborhood of entries by the same principles applying to the search of the graph homology.

With this reduction, [8, Theorem 3.10] is directly implied by Theorem 5.4 from Section 5.3 below.

Furthermore, the theory for matrices of arbitrary rank outlined in [9] can be interpreted as a non-linear generalization; furthermore, it indicates that matroid regression is also a viable tool for solving systems of equations carrying a structure of non-linear matroid.

Measuring Matrices and Phase Recognition As the low-rank matrix completion scenario indicates, the linear techniques can also be used if the signal x is in reality a matrix X, and each *row* of A is the vectorization a matrix Z_i of the same format, for example $Z_i = u_i v_i^*$, in which case $b_i = \mathrm{Tr}(XZ_i) + \varepsilon_i = v_i^* X u_i + \varepsilon$. Phase recognition is a special case of this example where $u_i = v_i$ for every i, and X- is a Hermitian rank one matrix. If one of u_i, v_i is always a standard basis vector e_i, and the other is $e_k \pm e_\ell$, this is a special subcase of the potentials or sparse sums scenario. If both are of the form $e_k - e_\ell$, the circuits correspond to the first syzygies of the rank one determinantal variety. In general, there is no easy way in which the circuits of the Z_i relate to those of u_i and v_i, but this is an interesting question to ask, in particular for the highly regular measurement designs employed in phase recognition.

The case where X is a symmetric, non-symmetric or partially symmetric tensor of higher degree, and where Z_i are outer products, can be seen as a generalization.

Low-Rank A In case the matrix A is of rank r and otherwise non-degenerate, matroid regression suggests an algorithm for inversion of A which includes inversions of $(r \times n)$ matrices only. Namely, any r rows will form a general regression circuit; if A is split into N/r disjoint row-blocks A_i of size r, then the estimator in Proposition 4.4 will be a weighted sum of estimates of the form $w^\top \mathbb{A}_i^{-1} b$, which is of lower complexity than inversion of A since matrix inversion scales with an exponent at least 2 in the size. This can be seen as an arbitrary rank generalization of the sample mean scenario, where the rank is 1.

5. Properties of the Matroid Regression Estimator

This section shows some key properties of the matroid regression estimator. Summarizing, we how that $\widehat{\gamma}(\mathcal{L})$ is the best linear unbiased estimator (BLUE) for any noise model, and the minimum variance unbiased estimator (MVUE) as well if the noise is Gaussian (homo- or heteroscedastic) - among all estimators that use only information in rows related to the particular system $\mathcal{L}$. Furthermore, we show a monotonicity result, showing that the variance of the estimator $\widehat{\gamma}(\mathcal{L})$ drops as $\mathcal{L}$ is enlarged. These results do not only show that the estimator $\widehat{\gamma}(\mathcal{L})$ is optimal, but also that Algorithm 3 computes a tight lower bound on the estimation error without actually estimating the evaluation $\langle w, x\rangle$, therefore provides a lower error bound for any method that is employed.

5.1. Monotonicity and Complexity-Accuracy-Tradeoff

The first important property of the estimator $\widehat{\gamma}(\mathcal{L})$ is being monotone with respect to inclusion of $\mathcal{L}$; that is, adding more circuits will only improve the estimator:

Theorem 5.1. *Let $\mathcal{L}, \mathcal{L}'$ be particular systems with $\mathcal{L} \subseteq \mathcal{L}'$. Then,* $\operatorname{Var}\left(\widehat{\gamma}(\mathcal{L}')\right) \leq \operatorname{Var}\left(\widehat{\gamma}(\mathcal{L})\right)$.

Proof. This follows from Theorem 4.1 (iii). □

Theorem 5.1 can be interpreted as a complexity-accuracy-tradeoff incurred by the amount of locality. More specifically, making the particular system $\mathcal{L}$ smaller will make Algorithm 1 run faster, but leads to an increased expected error in the estimate. Conversely, adding circuits and this enlarging $\mathcal{L}$ will make the estimate more accurate, but the algorithmic computation more expensive.

5.2. Optimality Amongst Linear Unbiased Estimators

The estimator $\widehat{\gamma}(\mathcal{L})$ has already been shown to be variance minimizing for choice of D in the particular system $\mathcal{L}$ in Theorem 4.1 (iii); we will make a similar statement relating it to using different entries of the vector b.

Definition 5.1. Let $\mathcal{L}$ be a particular/general system of divisors. The *support* of $\mathcal{L}$ is the inclusion-wise maximal set $I \subseteq [N]$ such that $\mathcal{L}$ contains all particular/general circuits contained in I. Conversely, for $I \subseteq [N]$, denote $\mathcal{L}(I) := \operatorname{aff}\{C \ : \ C \in \mathcal{C}_p(w|A), C \subseteq I\}$ or, equivalently, $\mathcal{L}(I) := C_p + \{C \ : \ C \in \mathcal{C}_c(w|A), C \subseteq I\}$ for some particular circuit $C_p \subseteq I$.

Theorem 5.2. *Let $I \subseteq [N]$, and let $\widehat{\gamma}'$ be any unbiased estimator for $\langle w, x\rangle$ linear in the entries $b_i, i \in I$, with coefficients depending only on A and w. Let $\mathcal{L} = \mathcal{L}(I)$. Then,* $\operatorname{Var}\left(\widehat{\gamma}(\mathcal{L})\right) \leq \operatorname{Var}\left(\widehat{\gamma}'\right)$.

Proof. This is implied by Proposition 4.1 which states that any estimator, linear in b_i and unbiased, is of the form $\widehat{\gamma}(D)$ for some divisor $D \in \mathcal{L}(I)$. The statement is then implied by Theorem 5.1. □

Theorem 5.2, together with the characterization of estimators linear in b in Proposition 4.1, implies that $\widehat{\gamma}$ is the best linear unbiased estimator (BLUE) for $\langle w, x\rangle$, and is an analogue of the Gauss-Markov theorem in our case.

5.3. Universal Variance Minimization

In this section, we will show that our estimator $\widehat{\gamma}$ is optimal in two further ways: first, for Gaussian noise, $\widehat{\gamma}$ is a sufficient and complete statistic, and thus the minimum variance unbiased estimator (MVUE); second for general centered noise, $\widehat{\gamma}$ has minimum variance among unbiased estimators independent of x and ε. This "noise optimality" implies that lower-variance estimators need to use additional information about either the unknown signal x or the distribution of the noise.

We first prove optimality for Gaussian noise:

Theorem 5.3. *Let $\mathcal{L}$ be a particular system. Assume that the noise ε is multivariate Gaussian. Then, the estimator $\widehat{\gamma}(\mathcal{L})$ is*

(i) a complete statistic with respect to the parameter $\langle w, x\rangle$ and observations $b_i, i \in \operatorname{supp}\mathcal{L}$.

(ii) a sufficient statistic with respect to the parameter $\langle w, x\rangle$ and observations $b_i, i \in \operatorname{supp}\mathcal{L}$.

(iii) the minimum variance unbiased estimator for the parameter $\langle w, x\rangle$ and observations $b_i, i \in \operatorname{supp}\mathcal{L}$.

Proof. (i) and (ii): Without loss of generality, we can assume that $\operatorname{supp}\mathcal{L}$ is all rows, otherwise, we remove the rows from A not contained in $\operatorname{supp}\mathcal{L}$.

Let $D_1, \dots, D_k$ be a basis for $\mathcal{L}$, and let Λ be the $(k \times N)$ matrix whose columns are the λ_{D_i}. Then, by definition, it holds that

$$\langle \Lambda, b\rangle = \langle w, x\rangle \cdot \mathbf{1} + \langle \Lambda, \varepsilon\rangle.$$

If we know that $\langle \Lambda, b\rangle$ is a complete and sufficient statistic for $\langle w, x\rangle$, we are done by virtue of the following argument: having reduced the original problem to linear regression, we observe that the BLUE of the latter is exactly $\widehat{\gamma}(D)$ with D minimizing $k(D, D)$, which is also known to be a complete and sufficient statistic and thus the MVUE, see e.g. [3, section 8.3, example 8.3] (for homoscedastic noise, the case of heteroscedastic noise follows from applying an appropriate linear transform). By Theorem 4.1, the estimator $\widehat{\gamma}(D)$ is the same as $\widehat{\gamma}(\mathcal{L})$, proving the statement.

We now prove the remaining **claim:** $\langle \Lambda, b\rangle$ is a complete and sufficient statistic for $\langle w, x\rangle$. To prove this, we observe that we can write x as an orthogonal decomposition $x = x_w + x_w^{\perp}$ where $x_w = \frac{x\langle w,x\rangle}{\langle w,x\rangle^2}$ is the orthogonal projection of x on $\operatorname{span} w$, thus $\langle w, x\rangle = \langle w, x_w\rangle$ and $\Lambda A x = \Lambda A x_w$. Since the columns of Λ are a basis of $\mathcal{L}$, the matrix ΛA acts, by definition, bijectively on x_w, which proves claim 1.

(iii) follows from (i) and (ii) via the Lehmann–Scheffé-Theorem. □

It is straightforward to extend the proof to other suitable members of the exponential family. By the Pitman-Koopman-Darmois theorem, it is unreasonable though to expect sufficiency for non-exponential distributions. However, we can prove a similar conclusion which relies on replacing sufficiency by universality with respect to the underlying signal:

Theorem 5.4. *Let $\mathcal{L}$ be a particular system. Let $\widehat{\gamma}'$ be an estimator for $\langle w, x\rangle$ which is unbiased for all choices of x, of the form $\widehat{\gamma}' = f(b,\ b \in \operatorname{supp}\mathcal{L})$ with some $f \in L^2(\mathbb{K}^n)$. Then,* $\operatorname{Var}\left(\widehat{\gamma}(\mathcal{L})\right) \geq \operatorname{Var}\left(\widehat{\gamma}'\right)$ *for any choice of noise ε.*

The proof of Theorem will be split in several statements. It will be immediately implied by Theorem 5.5 below, which states that an universally unbiased estimator always has the form $\widehat{\gamma}(\alpha)$, and Proposition 4.4.

Lemma 5.2. *Let $f \in L^2(\mathbb{K}^n)$. If $\mathbb{E}(f(X)) = 0$ for all random variables $X \in \mathbb{K}^n$ for which this expectation is finite, then $f = 0$.*

Proof. By definition, the statement is equivalent to: Let $f \in L^2(\mathbb{K}^n)$ be a smooth function. If

$$\langle f, p\rangle = \int_{\mathbb{K}^n} f(x)p(x)\,dx = 0$$

for all smooth functions $p : \mathbb{K}^n \to \mathbb{K}$ that fulfill $\langle p, 1\rangle = 1$ and $p(x) \geq 0$ for all $x \in \mathbb{K}^n$, then $f = 0$.

Now if $\langle f, p\rangle = 0$ for all smooth functions $p : \mathbb{K}^n \to \mathbb{R}$ that fulfill $\int_{\mathbb{K}^n} p(x)\ dx = 1$ and $p(x) \geq 0$, then $\langle f, g\rangle = 0$ for all functions $g \in L^2(\mathbb{K}^n)$, since the span of all such square-integrable p (which includes the simple functions) is dense in $L^2(\mathbb{K}^n)$. This implies $\langle f, f\rangle = \|f\|^2 = 0$, therefore $f = 0$. □

Lemma 5.3. *Let $f \in L^2(\mathbb{K}^n)$. If $\mathbb{E}(f(X)) = 0$ for all centered random variables $X \in \mathbb{K}^n$ for which this expectation is finite, then f is linear in X, i.e., of the form $f : x \mapsto \langle \lambda, x\rangle$ with $\lambda \in \mathbb{K}^n$.*

Proof. Denote $\phi_i : \mathbb{K}^n \to \mathbb{K}, x \mapsto x_i$, and denote

$$G = \{g \in L^2(\mathbb{K}^n)\ :\ \langle \phi_i, g\rangle < \infty \text{ for all } 1 \leq i \leq n\}.$$

Note that G is a $\mathbb{K}$-vector space. Further denote

$$H = G \cap \operatorname{span}\{p \in L^2(\mathbb{K}^n)\ :\ \langle 1, p\rangle = 1, \langle \phi_i, p \geq 0, p \text{ smooth}\}.$$

Since the square-integrable probability distributions span $L^2(\mathbb{K}^n)$, it follows that

$$H = \{g \in G\ :\ \langle \phi_i, g\rangle = 0 \text{ for all } 1 \leq i \leq n\}.$$

Therefore, H is the kernel of the linear map

$$\varphi : G \to \mathbb{K}^n, g \mapsto (\langle \phi_1, g\rangle, \ldots, \langle \phi_n, g\rangle).$$

This map is surjective, since e.g. all Gaussians are in G. By Lemma 5.2, $G^\perp = \{0\}$, therefore $H^\perp$ is contained in a vector space isomorphic to $\mathbb{K}^n$. Since the n-dimensional $\mathbb{K}$-vector space $(\mathbb{K}^n)^\vee$ is contained in $H^\perp$, it follows that $H^\perp = (\mathbb{K}^n)^\vee$. From this, the statement follows. □

An application yields the following statement:

Proposition 5.4. *Let $f \in L^2(\mathbb{K}^n)$, let $\beta \in \mathbb{K}^n$. If $\mathbb{E}(f(X)) = \langle \beta, \mathbb{E}(X)\rangle$ for all $\mathbb{K}^n$-valued random variables X for which the expectation $\mathbb{E}(f(X))$ is finite, then f is of the form $f : x \mapsto \langle \beta, x\rangle$.*

Proof. Applying Lemma 5.3 for the function $f - \langle \beta, .\rangle$ and the random variable $X - \mathbb{E}(X)$ yields that f is of the form

$$f : x \mapsto \langle \beta, \mathbb{E}(X)\rangle + \langle \lambda, x - \mathbb{E}(X)\rangle.$$

Since f is only a function of x and not of $\mathbb{E}(X)$ which can vary, the coefficient of $\mathbb{E}(X)$, which is equal to $\beta - \lambda$, must vanish, thus $\beta = \lambda$. Substituting yields the claim. □

Theorem 5.5. *Let $A \in \mathbb{K}^{N\times n}$ and $x \in \mathbb{K}^n$. Let ε be a centered and $\mathbb{K}^n$-valued random variable, let $b = Ax + \varepsilon$. For $c \in \operatorname{span} A$, let $\widehat{\gamma}$ be an estimator for $\langle w, x\rangle$ of the form $\widehat{\gamma} = f(b)$ with $f \in L^2(\mathbb{K}^n)$. If $\mathbb{E}(\widehat{\gamma}) = \langle w, x\rangle$ for all choices of x, and all choices of ε for which this expectation is finite, then $\widehat{\gamma}$ is of the form $\widehat{\gamma}(D)$, as described in Proposition 4.1.*

Proof. Since $w \in \operatorname{span} A$, there exists $\beta \in \mathbb{K}^N$ such that $\beta^\top AX = \langle w, x\rangle$. Linearity of expectation implies $\mathbb{E}(\langle \beta, b\rangle) = \langle w, x\rangle$. Taking $X = b$ and this β in Proposition 5.4 yields the claim. □

Acknowledgments

LT is supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no 247029- SDModels. This research was carried out at MFO, supported by FK's Oberwolfach Leibniz Fellowship.

References

[1] R. Barrett, M. W. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. van der Vorst. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. Society for Industrial and Applied Mathematics, 1994.

[2] F. R. Chung. *Spectral Graph Theory*. Number Nr. 92 in CBMS Regional Conference Series. American Mathematical Society, 1997.

[3] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman & Hall, 1st edition, 1974.

[4] T. A. Davis. *Direct Methods for Sparse Linear Systems*, volume 2 of *Fundamentals of Algorithms*. SIAM, 2006.

[5] W. Hackbusch. *Iterative Solution of Large Sparse Systems of Equations*. Springer, 1993.

[6] G. T. Herman. *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography*. Academic Press, 1980.

[7] A. C. Kak, M. Slaney, I. E. in Medicine, and B. Society. *Principles of Computerized Tomographic Imaging*. IEEE Engineering in Medicine and Biology Society, 1988.

[8] F. J. Király and L. Theran. Error-minimizing estimates and universal entry-wise error bounds for low-rank matrix completion. *Advances in Neural Information Processing Science 2013*, 2013.

[9] F. J. Király, L. Theran, R. Tomioka, and T. Uno. The algebraic combinatorial approach for low-rank matrix completion. Preprint, arXiv:1211.4116v4, 2012. URL `http://arxiv.org/abs/1211.4116`.

[10] Y. Saad. *Iterative Methods for Sparse Linear Systems: Second Edition*. Society for Industrial and Applied Mathematics, 2003.

[11] R. P. Tewarson. *Sparse Matrices*. Academic Press, 1973.