FINITE SAMPLE BOUNDS ON THE PERFORMANCE OF WEIGHTED LINEAR LEAST SQUARES IN SUB-GAUSSIAN CORRELATED NOISE

Michael Krikheli[1]

Faculty of Engineering,
Bar-Ilan University, 52900, Ramat-Gan
michael.krih@gmail.com

Amir Leshem [1]

Faculty of Engineering,
Bar-Ilan University, 52900, Ramat-Gan

ABSTRACT

In this paper we provide finite sample bounds on the performance of the weighted linear least squares estimator in sub-Gaussian martingale difference correlated noise. In contrast to standard performance analysis which uses bounds on the mean square error together with asymptotic normality, our bounds are based on concentration of measure. We extend previous results by analyzing the weighted least squares estimator and provide novel results in the case of correlated noise and heteroscedasticity. Using these bounds we obtain accurate bounds on the tail of the estimator. We show fast exponential convergence of the L^∞ probability of error. We analyze the fixed design setting. We use the results to analyze the performance of the weighted least squares estimator for the important problem of system identification. We show how to extend the results to different norms and state a theorem for the L^2 norm.

***Index Terms*—** Estimation; weighted least squares; non Gaussian; concentration bounds; finite sample; large deviations; confidence bounds; martingale difference sequence; system identification

1. INTRODUCTION

1.1. Related Work

Weighted linear least squares is one of the generalizations of the ordinary least squares. It has numerous applications in many fields, many times producing superior results to ordinary least squares. Notable examples would be [1, 2, 3, 4]. Standard analysis of estimators is asymptotic by nature. Cramer Rao bound (CRB) was utilized in many application, e.g. [5, 6]. The asymptotic properties of the least squares estimator were explored by various scientist both for the linear least squares model and the non-linear model. Notable works are [7, 8, 9, 10, 11, 12]. In these works large deviation results were given alongside confidence intervals analysis and convergence rate results. These works established the fact of the exponential convergence rate of least squares estimators under different conditions. Ridge regularized models were also analyzed by means of optimal convergence rate in [13].

The noise model differs in many applications of least squares and other optimization methods. Rather than the Gaussian model a Gaussian mixture is used in many applications. Notable examples of such applications are [14, 15, 16, 17]. In this work we consider sub-Gaussian noise, which is a general non-Gaussian noise framework. The Gaussian mixture model for instance is sub-Gaussian and our results are valid for this model. In the case of Gaussian noise least squares coincides with the maximum likelihood estimator. Still in many cases of interest least squares estimation is used in non-Gaussian noise as well for computation simplicity. Specifically the sub-Gaussian noise model is of special interest in many applications. In the past few years, the finite sample behavior of least squares problems has been studied in [18, 19, 20, 21, 22, 23]. These result show the finite sample behavior of ordinary least squares and regularized least squares under different noise conditions. Our result extends these results by analyzing the weighted least squares noise under very general noise conditions. While classical analysis usually assumes the i.i.d noise case, in many cases of interest the noise model used is not i.i.d but instead a martingale difference sequence model. This noise model is quite general and is used in various fields. For example [24, 25, 26]. These works utilize the martingale difference sequence model in different fields ranging from economics to control theory. The asymptotic properties of these models have been analyzed in various papers, for example [27, 28, 29, 26]. Recently, least squares under this general noise model was analyzed in [22, 23]. In this work we extend the results by analyzing the weighted least squares estimator. This estimator is used in many applications with heterogeneous data. We provide theoretical bounds on the uncertainty of the weighted least squares estimator in the presence of heteroscedasticity.

1.2. Contribution

In this paper we study the finite sample performance of the weighted least squares estimator. We assume an additive model with sub-Gaussian martingale difference noise and a fixed design matrix and heteroscedasticity in the noise random variables. Our bounds are given in the L^∞ norms for this general data model as well as L^2. We extend previous results [22, 23] to the weighted least squares case. This extension allows us to analyze heteroscedasticity in linear models and give finite sample results for the weighted least squares estimation technique in the sub-Gaussian martingale difference noise case. We show the connection between the number of required samples and the weights of the model. We demonstrate that in the presence of heteroscedasticity the weighted least squares outperforms the regular least squares estimator and the bounds capture this phenomenon. This is the first result for these types of models. We use the results to analyze the important signal processing problem of system identification in the presence of a bounded interfering signal passing through FIR channel.

2. PROBLEM FORMULATION

Consider a linear model with additive noise

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\theta}_0 + \boldsymbol{v} \tag{1}$$

[1]This work was supported by ISF grant 1644/18.

where $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$ is our output, $\boldsymbol{A} \in \mathbb{R}^{N\times p}$ is a known matrix with bounded random elements, $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is the parameter to estimate and $\boldsymbol{v} \in \mathbb{R}^{N\times 1}$ is a noise vector consisting of a zero mean sub-Gaussian white noise passing through a finite impulse response filter[12]. Important examples of $\boldsymbol{v}$ of this type are zero mean digital communication signals and Gaussian jammers. N indicates the number of samples used in the model and assumed to be larger than p.

Many real world noise models are sub-Gaussian; for instance, bounded constellation digital communication passing through a Rayleigh fading channel, finite Gaussian mixtures, any bounded random variable, and any combination of the above.

We write the assumptions detailed above in a mathematical form:

A1: $\boldsymbol{A} \in \mathbb{R}^{N\times p}$ is a fixed mixing matrix with $\alpha \doteq \max_{n,i} |a_{ni}|$.

A2: $E(v_n|F_{n-1}) = 0$, where F_{n-1} is a filtration and v_n, $n = 1 \dots N$ are the elements of $\boldsymbol{v}$.

A3: For every $1 \le j \le N$ v_j is sub-Gaussian with parameter δ_j^2.

The weighted least squares cost function is defined as:

$$J_0^N(\boldsymbol{\theta}, \boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{A\theta})^T \boldsymbol{W} (\boldsymbol{x} - \boldsymbol{A\theta}) \tag{2}$$

where $\boldsymbol{W} \in \mathbb{R}^{N\times N}$ is a diagonal matrix of weights, $\boldsymbol{W} \doteq \text{diag}\{w_1, \dots w_N\}$. Choosing a diagonal weight matrix is beneficial when our data has heteroscedasticity. This can happen when the conditions in which we acquire the data change, for example, if for a small subset of samples there is no interference. It would be optimal to weight these samples higher than the rest to achieve better performance. In this case $w_i = \delta_i$ where $\frac{1}{\delta_i}$ is the sub-Gaussian parameter of the noise plus interference of the model.

Given N samples the weighted least squares solution is given by

$$\hat{\boldsymbol{\theta}}_0^N = \left(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{x} = \left(\frac{1}{N}\sum_{n=1}^N w_n \boldsymbol{a}_n\boldsymbol{a}_n^T\right)^{-1}\frac{1}{N}\sum_{n=1}^N w_n \boldsymbol{a}_n^T x_n. \tag{3}$$

where $\boldsymbol{a}_n^T$, $n = 1 \dots N$ are the rows of $\boldsymbol{A}$ and x_n, $n = 1 \dots N$ are the data samples. $\hat{\boldsymbol{\theta}}_0^N$ is the optimum value of J_0^N. If the noise is zero mean then the estimator is unbiased.

We want to study the tail distribution of $\left\|\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right\|_\infty$ or more specifically we wish to bound the term

$$P\left(\left\|\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right\|_\infty > r\right) \le \varepsilon(N, r) \tag{4}$$

as a function of N and r. Furthermore, given N and r we want to calculate $\varepsilon(N, r)$ to achieve the above inequality. We analyze the case that the design matrix is fixed.

Throughout this paper we use the following mathematical notations:

Definition 2.1.

1. Let x be a random variable defined on the probability space (Ω, F, P) and denote $E(x)$ the expectation of x.
2. Let $\boldsymbol{B} \in \mathbb{R}^{p\times p}$ be a square matrix; we define the operators $\lambda_{max}(\boldsymbol{B})$ and $\lambda_{min}(\boldsymbol{B})$ to give the maximal and minimal eigenvalues of $\boldsymbol{B}$ respectively.
3. Let $\boldsymbol{C}$ be a matrix. The spectral norm for matrices is given by $\|\boldsymbol{C}\| \doteq \sqrt{\lambda_{max}(\boldsymbol{C}^T\boldsymbol{C})}$.
4. A random variable v with $E(v) = 0$ is called sub-Gaussian if its moment generating function exists and $E(\exp(sv)) \le \exp\left(\frac{s^2R^2}{2}\right)$ [30]. The minimal R^2 that satisfies this inequality is called the sub-Gaussian parameter of the random variable v and we say that v is sub-Gaussian with parameter R^2. Let $\boldsymbol{v} \in \mathbb{R}^N$ be a random vector. We call $\boldsymbol{v}$ sub-Gaussian with parameter R if each coordinate of $\boldsymbol{v}$ is sub-Gaussian with parameter R_i^2, $1 \le i \le N$ and $R^2 = \max_i\{R_i^2\}$.

[1]This is a special of a martingale difference noise. All the proofs remain valid in the general martingale difference noise case.

[2]For simplicity we only consider the real case. The complex case is similar with minor modifications.

3. MAIN RESULT

We are now ready to state and prove the main theorem.

Theorem 3.1. (Main Theorem)
Let $\boldsymbol{x}$ be defined as in (1) and assume A1-A3 hold. Given N, the number of samples and r the maximal error tolerated, the probability that the weighted least squares error is larger than r is bounded by

$$P\left(\left\|\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right\|_\infty > r\right) < \varepsilon(N, r) \tag{5}$$

where

$$\varepsilon(N, r) = p\exp\left(-\frac{N^2r^2\mu^2}{2\alpha^2\sum_{j=1}^N w_j^2\delta_j^2}\right) \tag{6}$$

and

$$\mu \doteq \lambda_{min}\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\right). \tag{7}$$

Proof. We wish to study the term

$$P\left(\left\|\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right\|_\infty > r\right). \tag{8}$$

In order to study this term, we will start by studying each coordinate separately. Then, we will use the union bound on the events that any of the coordinate error is larger than r. For each $1 \le i \le p$ separately consider

$$P\left(\left|\left(\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right)_i\right| > r\right). \tag{9}$$

We know that the weighted least squares solution is given by

$$\hat{\boldsymbol{\theta}}_0^N = \left(\frac{1}{N}\sum_{n=1}^N w_n\boldsymbol{a}_n\boldsymbol{a}_n^T\right)^{-1}\frac{1}{N}\sum_{n=1}^N w_n\boldsymbol{a}_n^T x_n. \tag{10}$$

$$\begin{aligned}
&\left|\left(\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right)_i\right| \\
&= \left(\left(\frac{1}{N}\sum_{n=1}^N w_n\boldsymbol{a}_n\boldsymbol{a}_n^T\right)^{-1}\frac{1}{N}\sum_{n=1}^N w_n\boldsymbol{a}_n^T v_n\right)_i \\
&\le \lambda_{max}\left(\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\right)^{-1}\right)\frac{1}{N}\sum_{n=1}^N w_n a_{ni} v_n \\
&= \frac{1}{\lambda_{min}\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\right)}\frac{1}{N}\sum_{n=1}^N w_n a_{ni} v_n.
\end{aligned} \tag{11}$$

We denote $\mu \doteq \lambda_{min}\left(\frac{1}{N}\boldsymbol{A^T W A}\right)$. Using this we can write

$$\begin{aligned} & P\left(\left|\left(\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right)_i\right| > r\right) \\ \leq \quad & P\left(\frac{1}{N}\sum_{n=1}^{N} w_n a_{ni} v_n > r\mu\right). \end{aligned} \tag{12}$$

In order to analyze this probability we use similar techniques as in [22, 23]. We will use the Laplace method. We start with bounding $E\left(\exp\left(s\sum_{n=1}^{N} w_n a_{ni} v_n\right)\right)$. We then use the achieved bound and Markov's inequality to bound the required probability.

$$\begin{aligned} & E\left(\exp\left(s\sum_{n=1}^{N} w_n a_{ni} v_n\right)\right) \\ = \quad & E\left(\exp\left(s\sum_{n=1}^{N-1} w_n a_{ni} v_n\right)\right) E\left(\exp\left(s w_N a_{Ni} v_N\right)|F_{N-1}\right) \\ \leq \quad & E\left(\exp\left(s\sum_{n=1}^{N-1} w_n a_{ni} v_n\right)\right) E\left(\exp\left(s w_N \alpha v_N\right)|F_{N-1}\right) \\ \leq \quad & E\left(\exp\left(s\sum_{n=1}^{N-1} w_n a_{ni} v_n\right)\right) \exp\left(\tfrac{s^2\alpha^2 w_N^2 \delta_N^2}{2}\right) \end{aligned} \tag{13}$$

The last inequality is due to the sub-Gaussianity of v_N. Iterating this procedure yields

$$E\left(\exp\left(s\sum_{n=1}^{N} w_n a_{ni} v_n\right)\right) \leq \exp\left(\frac{s^2\alpha^2}{2}\sum_{j=1}^{N} w_j^2\delta_j^2\right). \tag{14}$$

Looking now at the original equation we have

$$\begin{aligned} & P\left(\sum_{n=1}^{N} w_n a_{ni} v_n > Nr\mu\right) \\ \leq \quad & E\left(\exp\left(s\sum_{n=1}^{N} w_n a_{ni} v_n\right)\right)\exp\left(-sNr\mu\right) \\ \leq \quad & \exp\left(\tfrac{s^2\alpha^2}{2}\sum_{j=1}^{N} w_j^2\delta_j^2 - sNr\mu\right). \end{aligned} \tag{15}$$

The first inequality follows from Markov's inequality. The second inequality follows from equation (14). The inequality is true for all $s > 0$. Optimizing over s yields:

$$s = \frac{Nr\mu}{\alpha^2\sum_{j=1}^{N} w_j^2\delta_j^2}. \tag{16}$$

Substituting (16) into (15) we achieve

$$P\left(\sum_{n=1}^{N} w_n a_{ni} v_n > Nr\mu\right) \leq \exp\left(-\frac{N^2 r^2\mu^2}{2\alpha^2\sum_{j=1}^{N} w_j^2\delta_j^2}\right). \tag{17}$$

Using a union bound over the elements of the vector ensures that

$$P\left(\left\|\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right\|_\infty > r\right) \leq p\exp\left(-\frac{N^2 r^2\mu^2}{2\alpha^2\sum_{j=1}^{N} w_j^2\delta_j^2}\right) \tag{18}$$

and finishes the proof. □

Remark 3.2. If all v_n $n = 1 \ldots N$ have the same sub-Gaussian parameter δ^2 and if furthermore we choose $\boldsymbol{W} = \boldsymbol{I}$ then the main theorem bound can be written as

$$P\left(\left\|\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right\|_\infty > r\right) = p\exp\left(-\frac{Nr^2\lambda_{min}\left(\boldsymbol{A}^T\boldsymbol{A}\right)}{2\alpha^2\delta^2}\right) \tag{19}$$

which coincides with previous results analyzing the ordinary least squares estimator [22, 23].

While our analysis so far was for the L^∞ norm, the results easily translate to other norms using relationships between norms. We now state an equivalent theorem for the L^2 norm.

Theorem 3.3. *Let $\boldsymbol{x}$ be defined as in (1) and assume assumptions the assumptions about the model. Given N, the number of samples and r the maximal error tolerated, the probability that the weighted least squares error is larger than r is given by*

$$P\left(\left\|\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right\|_2 > r\right) \leq p\exp\left(-\frac{N^2 r^2\mu^2}{2\alpha^2 p\sum_{j=1}^{N} w_j^2\delta_j^2}\right). \tag{20}$$

4. SIMULATION RESULTS

In this section we analyze a problem of high importance in signal processing, i.e., system identification in the presence of interfering signal. In this case the fixed design matrix is a known training sequence of length N, $s_0, \ldots s_{N-1}$ that is transmitted through an unknown channel. The noise plus interference is modelled as a δ_i^2 sub-Gaussian martingale difference composed of another transmitter in the area as well as receiver noise. We model the case where the noise and interference parameters change during the training. This results in heteroscedasticity in the model. We show that the weighted least squares solution is better than the ordinary least squares solution and that the main theorem captures this well and gives easy to calculate performance bounds for this interesting case. As we now show the external interference which is a bounded communication signal passing through a linear time invariant channel is indeed a martingale difference sequence. The fixed design matrix is given by

$$\boldsymbol{A} = \begin{bmatrix} s_0 & 0 & 0 & \ldots & 0 \\ s_1 & s_0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ s_{p-1} & s_{p-2} & s_{p-3} & \ldots & s_0 \\ s_p & s_{p-1} & s_{p-2} & \ldots & s_1 \\ s_{p+1} & s_p & s_{p-1} & \ldots & s_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ s_{N-1} & s_{N-2} & s_{N-3} & \ldots & s_{N-p} \end{bmatrix} \tag{21}$$

where $s_0, \ldots s_{N-1}$ are random BPSK signals chosen in advance. $\boldsymbol{\theta}$ are the channel parameters to be estimated using least squares. The mathematical setting is given by

$$x_n = \sum_{t=0}^{p-1} s_{n-t}\theta_t + v_n \tag{22}$$

where we define $s_i \doteq 0 \quad \forall i < 0$. This equation can be written as

$$\boldsymbol{x} = \boldsymbol{A\theta} + \boldsymbol{v} \tag{23}$$

where $\boldsymbol{A}$ is defined in (21). The noise vector $\boldsymbol{v} \doteq (v_0, \ldots, v_{N-1})^T$ can be modelled as:

$$v_n = \sum_{i=0}^{k} h_i j_{n-i} + w_n \tag{24}$$

where j_n is i.i.d zero mean bounded signal for example a BPSK signal. This can happen for example when estimating the channel in a CDMA sequence, when the interference is composed of another CDMA signal. [3] We denote by η the bound for j_n, i.e. $P(j_n \leq \eta) = 1$. We also assume that h_i is an unknown system modelled as a sub-Gaussian random variable. It is easy to see that this noise which is a typical interference scenario is indeed a zero mean martingale difference sequence. Under this model 10% of the samples are sub-Gaussian random variables δ_1^2 and δ_2^2 respectively. 90% of the samples are random variables with sub-Gaussian parameters given by $4\delta_1^2$ and $4\delta_2^2$. This results in heteroscedasticity in the data model. We use the fact that $j_n \leq \eta$ and the fact that a linear combination of sub-Gaussian random variables is sub-Gaussian [30]. We can conclude that the noise is sub-Gaussian when for 10% the parameter is $\delta_1^2\eta^2 + \delta_2^2$ and for 90% of the samples the parameter is $4\delta_1^2\eta^2 + 4\delta_2^2$. We can use the main theorem with these parameters to calculate the performance for different values of N. Figure 1 depicts the relationship between the main theorem bound and the simulation. It shows that the performance of the bound is similar to the simulation performance and that the bound can be used to reliably upper bound the performance of the estimator given the model parameters. Figure 2 demonstrates the advantages of the weighted least squares estimator over the ordinary least squares estimator in the presence of heteroscedasticity. We show that both the simulation results and the bounds for the weighted least squares estimator outperform the ordinary least squares estimator. This shows that the bounds in the main theorem correctly captures the advantages of the weighted least squares approach.

5. CONCLUDING REMARKS

In this paper we examined the finite sample performance of the L^∞ error of the weighted linear least squares estimator martingale difference sequence with heteroscedasticity as our data model. We showed a very fast convergence of the L^∞ error probability as a function of the required performance and the number of samples. We extended previous results by analyzing the weighted least squares. This allows us to analyze problems with changing conditions and heteroscedasticity. These settings could not be analyzed with previous results. While our analysis was conducted for the L^∞ norm, this doesn't limit the scope of our methods as relationships between norms can be utilized to bound the error vector under different norms. We demonstrate this by stating an equivalent theorem for the L^2 norm. We show that the in the presence of heteroscedasticity the weighted least squares estimator outperforms the regular least squares estimator. We show that this relationship between holds for the bounds depicted in the main theorem of this paper and the main theorem of [22].

[3] Similar analysis is relevant for high range resolution (HRR) estimation of target parameters in the presence of temporally correlated jammer. See [31] for example.

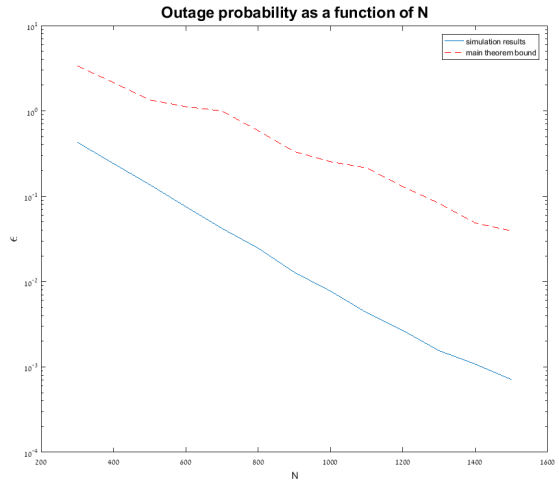

Fig. 1. Finite sample performance analysis of the weighted least squares estimator for the system identification problem with sub-Gaussian martingale difference noise plus interference. The graph shows simulation results and main theorem bounds for parameters $r = 0.08$ and $p = 8$.

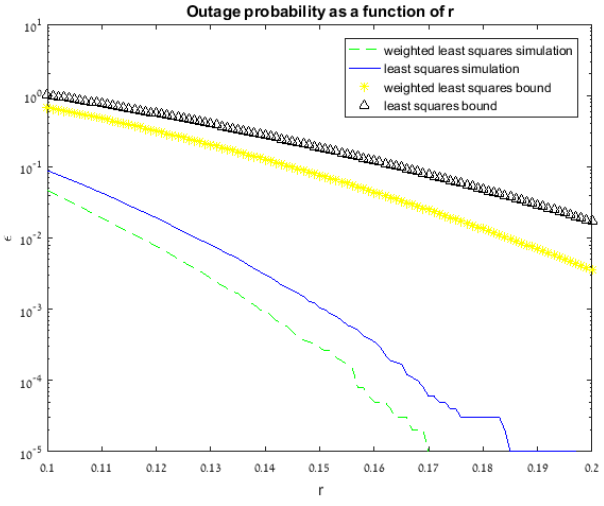

Fig. 2. Performance analysis of weighted least squares and linear least squares estimators with sub-Gaussian martingale difference noise. The graph shows simulation results and theoretical bounds for both estimators with parameters $N = 300$ and $p = 4$.

6. REFERENCES

[1] X. Zhang and X. Wu, "Image interpolation by adaptive 2-D autoregressive modeling and soft-decision estimation," *IEEE Transactions on Image Processing*, vol. 17, no. 6, pp. 887–896, June 2008.

[2] K. W. Hung and W. C. Siu, "Robust soft-decision interpolation using weighted least squares," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1061–1069, March 2012.

[3] H. C. So and L. Lin, "Linear least squares approach for accurate received signal strength based source localization," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 4035–4040, Aug 2011.

[4] Jelle Veraart, Jan Sijbers, Stefan Sunaert, Alexander Leemans, and Ben Jeurissen, "Weighted linear least squares estimation of diffusion MRI parameters: strengths, limitations, and pitfalls," *Neuroimage*, vol. 81, pp. 335–346, 2013.

[5] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood and Cramer-Rao bound," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, Apr 1988, pp. 2296–2299 vol.4.

[6] A. Leshem and A. J. van der Veen, "Direction-of-arrival estimation for constant modulus signals," *IEEE Transactions on Signal Processing*, vol. 47, no. 11, pp. 3125–3129, Nov 1999.

[7] Leon J Gleser, "On the asymptotic theory of fixed-size sequential confidence bounds for linear regression parameters," *The Annals of Mathematical Statistics*, pp. 463–467, 1965.

[8] B.L.S Prakasa Rao, "On the exponential rate of convergence of the least squares estimator in the nonlinear regression model with Gaussian errors," *Statistics & probability letters*, vol. 2, no. 3, pp. 139–142, 1984.

[9] B.L.S Prakasa Rao, "The rate of convergence of the least squares estimator in a non-linear regression model with dependent errors," *Journal of Multivariate Analysis*, vol. 14, no. 3, pp. 315–322, 1984.

[10] Chien-Fu Wu, "Asymptotic theory of nonlinear least squares estimation," *The Annals of Statistics*, pp. 501–513, 1981.

[11] Arthur Sieders and Kacha Dzhaparidze, "A large deviation result for parameter estimators and its application to nonlinear regression analysis," *The Annals of Statistics*, pp. 1031–1049, 1987.

[12] Hu Shuhe, "A large deviation result for the least squares estimators in nonlinear regression," *Stochastic processes and their applications*, vol. 47, no. 2, pp. 345–352, 1993.

[13] Andrea Caponnetto and Ernesto De Vito, "Optimal rates for the regularized least-squares algorithm," *Foundations of Computational Mathematics*, vol. 7, no. 3, pp. 331–368, 2007.

[14] S. Banerjee and M. Agrawal, "On the performance of underwater communication system in noise with Gaussian mixture statistics," in *2014 Twentieth National Conference on Communications (NCC)*, Feb 2014, pp. 1–6.

[15] J. Tan, D. Baron, and L. Dai, "Wiener filters in Gaussian mixture signal estimation with ℓ_∞ -norm error," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6626–6635, Oct 2014.

[16] Vimal Bhatia and Bernard Mulgrew, "Non-parametric likelihood based channel estimator for Gaussian mixture noise," *Signal Processing*, vol. 87, no. 11, pp. 2569–2586, 2007.

[17] Xiaodong Wang and H. V. Poor, "Robust multiuser detection in non-Gaussian channels," *IEEE Transactions on Signal Processing*, vol. 47, no. 2, pp. 289–305, Feb 1999.

[18] Roberto Imbuzeiro Oliveira, "The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties," *arXiv preprint arXiv:1312.2903*, 2013.

[19] Daniel Hsu, Sham M Kakade, and Tong Zhang, "Random design analysis of ridge regression," *Foundations of Computational Mathematics*, vol. 14, no. 3, pp. 569–600, 2014.

[20] Jean-Yves Audibert and Olivier Catoni, "Robust linear regression through PAC-Bayesian truncation," *Preprint, URL http://arxiv. org/abs/1010.0072*, vol. 38, pp. 60, 2010.

[21] Jean-Yves Audibert and Olivier Catoni, "Robust linear least squares regression," *The Annals of Statistics*, pp. 2766–2794, 2011.

[22] Michael Krikheli and Amir Leshem, "Finite sample performance of linear least squares estimators under sub-Gaussian martingale difference noise," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4444–4448.

[23] Michael Krikheli and Amir Leshem, "Finite sample performance of linear least squares estimation," *arXiv preprint arXiv:1810.06380*, 2018.

[24] Robert F Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation," *Econometrica: Journal of the Econometric Society*, pp. 987–1007, 1982.

[25] Tze Leung Lai and Ching Zong Wei, "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *The Annals of Statistics*, pp. 154–166, 1982.

[26] Walter Krämer, "Finite sample efficiency of ordinary least squares in the linear regression model with autocorrelated errors," *Journal of the American Statistical Association*, vol. 75, no. 372, pp. 1005–1009, 1980.

[27] TL Lai and CZ Wei, "Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters," *Journal of multivariate analysis*, vol. 13, no. 1, pp. 1–23, 1983.

[28] Paul I Nelson, "A note on strong consistency of least squares estimators in regression models with martingale difference errors," *The Annals of Statistics*, pp. 1057–1064, 1980.

[29] Norbert Christopeit and Kurt Helmes, "Strong consistency of least squares estimators in linear regression models," *The Annals of Statistics*, pp. 778–788, 1980.

[30] Roman Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, p. 210268, Cambridge University Press, 2012.

[31] Jian Li, Guoqing Liu, Nanzhi Jiang, and Petre Stoica, "Airborne phased array radar: clutter and jamming suppression and moving target detection and feature extraction," in *Sensor Array and Multichannel Signal Processing Workshop. 2000. Proceedings of the 2000 IEEE*. IEEE, 2000, pp. 240–244.