Approximating the Row-Wise Weighted Total Least Squares Linear Regression Solution

Cole Plepel[1]
[1] **Illinois Mathematics and Science Academy**

ABSTRACT

Motivated by applications as a kernel of nonlinear regression algorithms, the row-wise weighted total least squares regression problem is examined to find a consistent and accurate estimator. Specifically, the estimator will have a time complexity linear in the number of observations and a space complexity constant in the same value, as the number of observations can be quite large in many modern applications, often many orders of magnitude larger than the number of input and output features. Further, to accommodate large data sets, an algorithm is sought to update an intermediate representation from each observation, allowing for parallelization of the necessary computation. Four related algorithms are proposed, based on approximating the noncentral second moment of the underlying data by a weighted mean, requiring only linear time in the number of observations. Experimental findings show the proposed algorithm to be competitive with existing methods intended to solve other variants of the Total Least Squares problem. Directions for continued iteration and further investigation are proposed.

ACKNOWLEDGEMENTS

This inquiry has been pursued as an SIR project advised by Dr. Evan Glazer, President of the Illinois Mathematics and Science Academy. Thanks also to Dr. Amhet Cetin and Dr. Mesrob Ohannessian of the University of Illinois Chicago and Dr. Aritra Dutta of the University of Southern Denmark for their feedback and advice.

INTRODUCTION

Motivation

- Linear regression is useful in economics, scientific modeling, signal denoising, and, recently, machine learning.
- Neural Networks are usually trained with Stochastic Gradient Descent, but this is very slow.
- Extreme Learning Machines (ELMs) (Huang *et al*.) are a recent method of training neural networks very quickly, but it only works for Single-Layer Feed-Forward neural networks, a very restrictive class.
- Extensions to deep neural networks exist, such as Multi-Layer ELM and Deep ELM, but they ignore second-order information, allowing noise to build up over layers. A method of considering which directions are easier or harder to modify may help improve these methods.

Background

- Ordinary Least Squares (OLS) is one common method of linear regression, which minimizes the sum of the squared distances of the points from the regression in the y-direction only.
  - This is equivalent to assuming data is exact in the x-direction, but noisy in the y-direction, and the amount of noise is the same for each observation, but this is often an unreasonable assumption.
- Total Least Squares (TLS) accounts for errors in input variables by minimizing the sum of the squared distances of each point to the regression perpendicular to the regression.
  - This is equivalent to assuming isotropic noise distributions and the same amount of noise for each observation. Noisy often has internal correlations this doesn't account for, however.
- Generalized Total Least Squares (GTLS) allows the noise in the input and output variables to be correlated. This means the noise is assumed to be homoscedastic (the same), but not isotropic (uncorrelated) for all samples.
- Row-Wise Weighted Total Least Squares (RW-TLS) expands on this, allowing each sample to have a different noise distribution. This is called heteroskedasticity, and it provides a very general model, applicable to many situations.

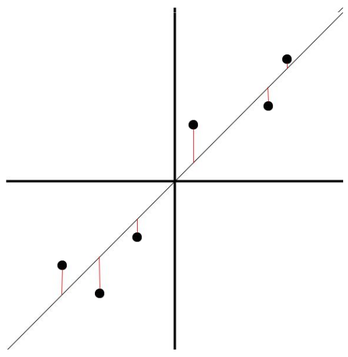


Figure 1. OLS Regression Error. OLS minimizes the sum of the squares of the red distances.

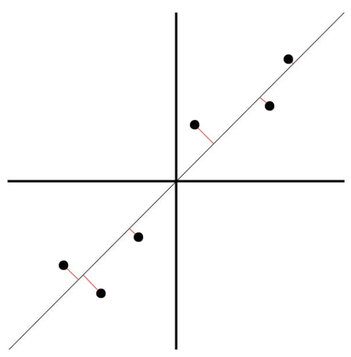


Figure 2. TLS Regression Error. TLS minimizes the sum of the squares of the red distances.

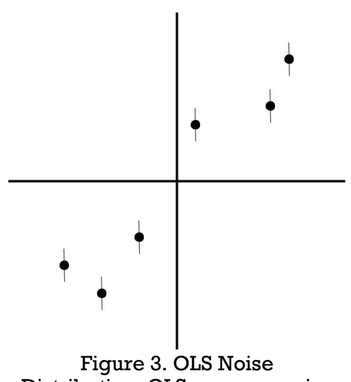


Figure 3. OLS Noise Distribution. OLS assumes noise exists only in the y direction, and that the variance is constant.

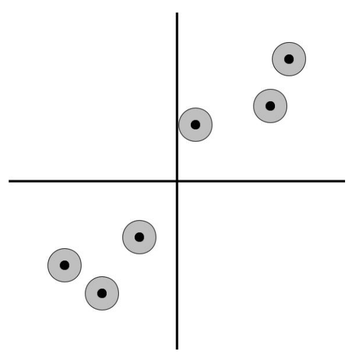


Figure 4. TLS Noise Distribution. TLS assumes noise is isotropic with constant variance across data points.

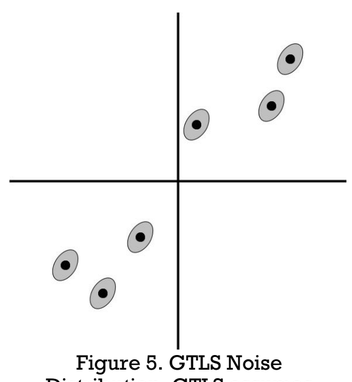


Figure 5. GTLS Noise Distribution. GTLS assumes noise follows the same distribution for each data point.

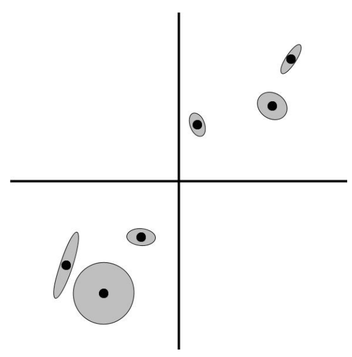


Figure 6. RW-TLS Noise Distribution. RW-TLS allows the noise distributions to vary arbitrarily between data points.

CRITERIA

- Time complexity: linear in number of observations, ideally parallelizable
  - The number of observations can be far greater than the number of features in many applications
- Statistical consistency: convergence to the underlying regression
- Experimental utility: data reconstruction and recovery of weight matrix

PROPOSED APPROACH

- The ideal solution could be recovered from the covariance matrix of the underlying data through eigen-decomposition.
- Each observation provides an estimate of this covariance matrix.
- Estimate the underlying covariance matrix as the mean of the estimates, weighted by their respective precision tensors.
- Alternatively, use a scalar, such as the inverse of the spectral radius, determinant, or trace of the noise covariance matrix as the weight

Algorithm 3 Precision Method

Require: $\{x_n\}_{n=1}^N \in \mathbb{R}^F$, $\{y_n\}_{n=1}^N \in \mathbb{R}^K$, $\{\Sigma_n^{(x)}\}_{n=1}^N \in \mathbb{R}^{F\times F}$, $\{\Sigma_n^{(y)}\}_{n=1}^N \in \mathbb{R}^{K\times K}$, $\{\Sigma_n^{(xy)}\}_{n=1}^N \in \mathbb{R}^{F\times K}$

$Z \leftarrow 0_{(F+K)\times(F+K)}$
$P \leftarrow 0_{(F+K)\times(F+K)\times(F+K)\times(F+K)}$
for $n \leftarrow 1..N$ **do**
  $c \leftarrow \begin{bmatrix} x_n \\ y_n \end{bmatrix}$
  $\Sigma \leftarrow \begin{bmatrix} \Sigma_n^{(x)} & \Sigma_n^{(xy)} \\ \Sigma_n^{(xy)\top} & \Sigma_n^{(y)} \end{bmatrix}$
  $M_{ij} \leftarrow c_i c_j + \Sigma_{ij}$
  $V_{ijkl} \leftarrow c_i c_k \Sigma_{jl} + c_i c_l \Sigma_{jk} + c_j c_k \Sigma_{il} + c_j c_l \Sigma_{ik} + \Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk}$
  $Z_{ij} \leftarrow Z_{ij} + V_{ijkl}^{-1} M_{kl}$
  $P \leftarrow P + V^{-1}$
end for
$E_{ij} \leftarrow P_{ijkl}^{-1} Z_{kl}$
$Q, \Lambda \leftarrow \text{EIGEN}(E)$
return $(Q_{:F,:F}^\top)^{-1} Q_{:F,-K:}^\top$

Algorithm 4 Radius Method

Require: $\{x_n\}_{n=1}^N \in \mathbb{R}^F$, $\{y_n\}_{n=1}^N \in \mathbb{R}^K$, $\{\Sigma_n^{(x)}\}_{n=1}^N \in \mathbb{R}^{F\times F}$, $\{\Sigma_n^{(y)}\}_{n=1}^N \in \mathbb{R}^{K\times K}$, $\{\Sigma_n^{(xy)}\}_{n=1}^N \in \mathbb{R}^{F\times K}$

$Z \leftarrow 0_{(F+K)\times(F+K)}$
$w \leftarrow 0$
for $n \leftarrow 1..N$ **do**
  $c \leftarrow \begin{bmatrix} x_n \\ y_n \end{bmatrix}$
  $\Sigma \leftarrow \begin{bmatrix} \Sigma_n^{(x)} & \Sigma_n^{(xy)} \\ \Sigma_n^{(xy)\top} & \Sigma_n^{(y)} \end{bmatrix}$
  $M_{ij} \leftarrow c_i c_j + \Sigma_{ij}$
  $v = \rho(\Sigma)$
  $Z \leftarrow Z + V_{-1}^{-1} M_{A1}$
  $w \leftarrow w + v$
end for
$E_{ij} \leftarrow \frac{1}{w} Z$
$Q, \Lambda \leftarrow \text{EIGEN}(E)$
return $(Q_{:F,:F}^\top)^{-1} Q_{:F,-K:}^\top$

Algorithm 5 Determinant Method

Require: $\{x_n\}_{n=1}^N \in \mathbb{R}^F$, $\{y_n\}_{n=1}^N \in \mathbb{R}^K$, $\{\Sigma_n^{(x)}\}_{n=1}^N \in \mathbb{R}^{F\times F}$, $\{\Sigma_n^{(y)}\}_{n=1}^N \in \mathbb{R}^{K\times K}$, $\{\Sigma_n^{(xy)}\}_{n=1}^N \in \mathbb{R}^{F\times K}$

$Z \leftarrow 0_{(F+K)\times(F+K)}$
$w \leftarrow 0$
for $n \leftarrow 1..N$ **do**
  $c \leftarrow \begin{bmatrix} x_n \\ y_n \end{bmatrix}$
  $\Sigma \leftarrow \begin{bmatrix} \Sigma_n^{(x)} & \Sigma_n^{(xy)} \\ \Sigma_n^{(xy)\top} & \Sigma_n^{(y)} \end{bmatrix}$
  $M_{ij} \leftarrow c_i c_j + \Sigma_{ij}$
  $v = \det(\Sigma)$
  $Z \leftarrow Z + V_{-1}^{-1} M_{A1}$
  $w \leftarrow w + v$
end for
$E_{ij} \leftarrow \frac{1}{w} Z$
$Q, \Lambda \leftarrow \text{EIGEN}(E)$
return $(Q_{:F,:F}^\top)^{-1} Q_{:F,-K:}^\top$

Algorithm 6 Trace Method

Require: $\{x_n\}_{n=1}^N \in \mathbb{R}^F$, $\{y_n\}_{n=1}^N \in \mathbb{R}^K$, $\{\Sigma_n^{(x)}\}_{n=1}^N \in \mathbb{R}^{F\times F}$, $\{\Sigma_n^{(y)}\}_{n=1}^N \in \mathbb{R}^{K\times K}$, $\{\Sigma_n^{(xy)}\}_{n=1}^N \in \mathbb{R}^{F\times K}$

$Z \leftarrow 0_{(F+K)\times(F+K)}$
$w \leftarrow 0$
for $n \leftarrow 1..N$ **do**
  $c \leftarrow \begin{bmatrix} x_n \\ y_n \end{bmatrix}$
  $\Sigma \leftarrow \begin{bmatrix} \Sigma_n^{(x)} & \Sigma_n^{(xy)} \\ \Sigma_n^{(xy)\top} & \Sigma_n^{(y)} \end{bmatrix}$
  $M_{ij} \leftarrow c_i c_j + \Sigma_{ij}$
  $v = \text{Tr}(\Sigma)$
  $Z \leftarrow Z + V_{-1}^{-1} M_{A1}$
  $w \leftarrow w + v$
end for
$E_{ij} \leftarrow \frac{1}{w} Z$
$Q, \Lambda \leftarrow \text{EIGEN}(E)$
return $(Q_{:F,:F}^\top)^{-1} Q_{:F,-K:}^\top$

METHODOLOGY

- 100 datasets were generated with 5 input features and 3 output features, following a normal distribution, with each of 100, 1,000, 10,000, and 100,000 samples.
- Isotropic, homoscedastic, and heteroskedastic noise was added to each dataset.
- TLS, GTLS, and each algorithm were used to fit each dataset.
- The mean and standard error of the weight recovery error and data reconstruction error were reported

Algorithm 1 TLS Algorithm

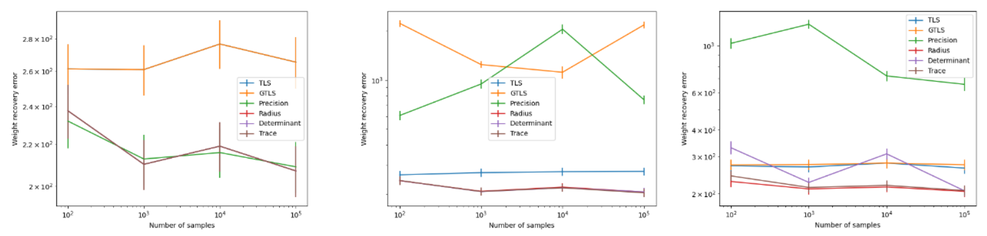
Require: $X \in \mathbb{R}^{N\times F}$, $Y \in \mathbb{R}^{N\times K}$

$C \leftarrow \begin{bmatrix} X \\ Y \end{bmatrix}$
$U, \Sigma, V^\top \leftarrow \text{SVD}(C)$
return $(V_{:F,:F}^\top)^{-1} V_{:F,-K:}^\top$

Algorithm 2 GTLS Algorithm

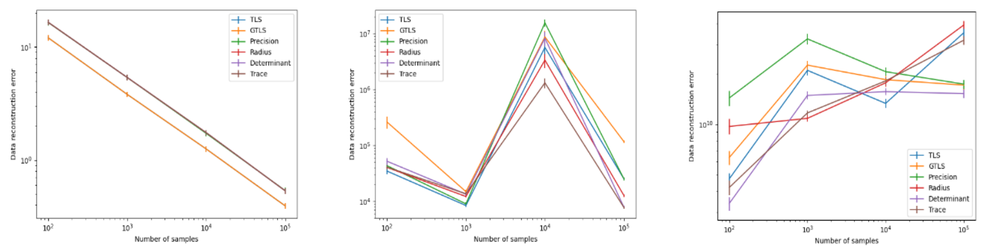
Require: $X \in \mathbb{R}^{N\times F}$, $Y \in \mathbb{R}^{N\times K}$, $\Sigma_X \in \mathbb{R}^{F\times F}$, $\Sigma_Y \in \mathbb{R}^{K\times K}$

$C \leftarrow \begin{bmatrix} X\Sigma_X^{-\frac{1}{2}} \\ Y\Sigma_Y^{-\frac{1}{2}} \end{bmatrix}$
$U, \Sigma, V^\top \leftarrow \text{SVD}(C)$
return $\Sigma_X^{-\frac{1}{2}} (V_{:F,:F}^\top)^{-1} V_{:F,-K:}^\top \Sigma_Y^{\frac{1}{2}}$



(a) Isotropic noise (b) Homoskedastic noise (c) Heteroskedastic noise

Figure 1: Weight recovery error by algorithm and number of samples for (a) isotropic, (b) homoskedastic, and (c) heteroskedastic noise conditions. Error bars represent standard errors of the reported means, across 100 trials.



(a) Isotropic noise (b) Homoskedastic noise (c) Heteroskedastic noise

Figure 2: Data reconstruction error by algorithm and number of samples for (a) isotropic, (b) homoskedastic, and (c) heteroskedastic noise conditions. Error bars represent standard errors of the reported means, across 100 trials.

DISCUSSION

- Experimental results are promising for the proposed algorithms.
- Routes for further improvement exist. For instance:
- The mapping from *E* to the estimated weight matrix is highly nonlinear. Accounting for this nonlinearity may allow for a more statistically efficient estimator.
- Other weights on the influence of each *M* matrix on the estimate of the covariance matrix of the underlying data may yield further improvements in accuracy or computational efficiency.

CONCLUSION

An effective and efficient algorithm for the Row-Wise Weighted Total Least Squares linear regression problem would have applications in various fields, such as economics, image processing, statistical modelling, and machine learning. Approximating the covariance matrix of the underlying data appears to provide an effective basis for such a solution. The proposed algorithms perform comparably with existing methods, and variants of the initial algorithm improve on the experimental accuracy of the algorithm, outperforming existing methods on the sample datasets. Routes exist to further improve on the proposed algorithms. Additional routes exist which may further improve the accuracy and efficiency of the proposed algorithm.

REFERENCES

[1] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," Neurocomputing, vol. 70, no. 1, pp. 489–501, 2006, Neural Networks. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231206000385
[2] S. Ding, N. Zhang, X. Xu, L. Guo, and J. Zhang, "Deep extreme learning machine and its application in EEG classification," Mathematical Problems in Engineering, vol. 2015, May 2015.
[3] L. Kasun, H. Zhou, G.-B. Huang, and C.-M. Vong, "Representational learning with ELMs for big data," IEEE Intelligent Systems, vol. 28, no. 6, pp. 31–34, Nov. 2013.
[4] J. Tang, C. Deng, G.-B. Huang, and J. Hou, "A fast learning algorithm for multi-layer extreme learning machine," in 2014 IEEE International Conference on Image Processing (ICIP), Oct. 2014, pp. 175–178.
[5] J. Zhao, Z. Wang, and F. Cao, "Extreme learning machine with errors in variables," World Wide Web, vol. 17, no. 5, pp. 1205–1216, Sep. 2014, copyright - Springer Science+Business Media New York 2014; Last updated - 2015-02-18.
[6] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., May 2015. [Online]. Available: https://arxiv.org/abs/1412.6980
[7] M. Zeiler, "Adadelta: An adaptive learning rate method," arXiv: Machine Learning, Dec. 2012. [Online]. Available: https://arxiv.org/abs/1212.5701
[8] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," Journal of Machine Learning Research, vol. 12, p. 2121–2159, Jul. 2011.
[9] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," arXiv: Computer Vision and Pattern Recognition, vol. abs/1912.04958, 2019. [Online]. Available: http://arxiv.org/abs/1912.04958
[10] S. Becker and Y. Lecun, "Improving the convergence of back- propagation learning with second-order methods," in Proceedings of the 1988 Connectionist Models Summer School, D. Touretzky, G. Hinton, and T. Sejnowski, Eds. Morgan Kaufmann, Jan. 1989, pp. 29–37.
[11] I. Markovsky and S. Van Huffel, "Overview of total least-squares methods," Signal Processing, vol. 87, no. 10, pp. 2283–2302, 2007, special Section: Total Least Squares and Errors-in-Variables Modeling. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168407001405
[12] N. Srebro and T. Jaakkola, "Weighted low-rank approximations," in Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003). AAAI Press, 2003, pp. 720–727.
[13] D. Malioutov and N. Slavov, "Convex total least squares," in Proceedings of the 31st International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, no. 2. Bejing, China: PMLR, Jun. 2014, pp. 109–117. [Online]. Available: https://proceedings.mlr.press/v32/malioutov14.html
[14] A. Dutta, "Weighted low-rank approximation of matrices: Some analytical and numerical aspects," 2016. [Online]. Available: https://stars.library.ucf.edu/etd/5631
[15] L. Isserlis, "On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables," Biometrika, vol. 12, no. 1/2, pp. 134–139, 1918. [Online]. Available: http://www.jstor.org/stable/2331932
[16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
[17] A. Dutta, X. Li, and P. Richtarik, "Weighted low-rank approximation of matrices and background modeling," arXiv: Computer Vision and Pattern Recognition, Apr. 2018. [Online]. Available: https://arxiv.org/abs/1804.06252
[18] D. Fasino and A. Fazzi, "A Gauss-Netwon iteration for total least squares problems," BIT Numerical Mathematics, vol. 58, no. 2, pp. 281–299, 2018.
[19] Y. Tsaregorodtsev, "Asymptotic normality of element-wise weighted total least squares estimator in a multivariate error-in-variables model," arXiv: Statistical Theory, 2017. [Online]. Available: https://arxiv.org/abs/1702.00842
[20] I. Markovsky, M. Rastello, A. Premoli, A. Kukush, and S. Van Huffel, "The element-wise weighted total least squares problem," Computational Statistics & Data Analysis, vol. 51, pp. 181–209, Jan. 2005.
[21] S. Shkylar, "Conditions for the consistency of the total least squares estimator in an error-in-variables linear regression model," Theory of Probability and Mathematical Statistics, vol. 83, pp. 175–190, Jan. 2011.
[22] S. Van Huffel and J. Vandewalle, "Analysis and properties of the generalized total least squares problem AX ≈ B when some or all columns in A are subject to error," SIAM Journal on Matrix Analysis and Applications, vol. 10, no. 3, pp. 294–315, Jul. 1989.