Technical Note: Line-Fit Estimation

Paul F. Roysdon, Ph.D.

I. Introduction

In this Technical Note we derive the theory for a simple line-fitting problem, then demonstrate the theory with an example. The goal is to develop a method for outlier accommodation, wherein the approach is seeks to use the minimum number of measurements to fit a model. Eventually this method will be applied to a full-nonlinear GPS-INS sliding-window problem.

II. Theory

A. Line-Fit

Consider a simple line fitting problem, where the true line is represented by

$$\mathbf{y} = a\boldsymbol{x} + b, \tag{1}$$

for $\boldsymbol{x}, \mathbf{y} \in \mathbb{R}^{n\times 1}$.

Given measurements $\tilde{\mathbf{z}}_i$ for $i = 1\ldots n$, eqn. (1) is modified to be

$$\tilde{\mathbf{z}}_i = a\boldsymbol{x}_i + b + \boldsymbol{\eta}_i, \tag{2}$$

where $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

A method for solving the line-fit problem is Bayesian Mean Square Error (BMSE) [1], such that

$$\text{BMSE} = \sum_{i=1}^{n} (\tilde{\mathbf{z}}_i - a\boldsymbol{x}_i - b)^2. \tag{3}$$

Linearization of eqn. (3) requires partial derivatives of a and b, such that

$$\begin{aligned}
\frac{\partial}{\partial a}\text{BMSE} &= 2\sum_{i=1}^{n}(\tilde{\mathbf{z}}_i - a\boldsymbol{x}_i - b)(-\boldsymbol{x}_i) = 0\\
&= -\sum_{i=1}^{n}\tilde{\mathbf{z}}_i\boldsymbol{x}_i + a\sum_{i=1}^{n}\boldsymbol{x}_i^2 + b\sum_{i=1}^{n}\boldsymbol{x}_i = 0, \qquad (4)\\
\frac{\partial}{\partial b}\text{BMSE} &= 2\sum_{i=1}^{n}(\tilde{\mathbf{z}}_i - a\boldsymbol{x}_i - b)(-1) = 0\\
&= -\sum_{i=1}^{n}\tilde{\mathbf{z}}_i + a\sum_{i=1}^{n}\boldsymbol{x}_i + bn = 0. \qquad (5)
\end{aligned}$$

Rearranging terms in eqns. (4) and (5) produces

$$a\sum_{i=1}^{n}\boldsymbol{x}_i^2 + b\sum_{i=1}^{n}\boldsymbol{x}_i = \sum_{i=1}^{n}\tilde{\mathbf{z}}_i\boldsymbol{x}_i \tag{6}$$

$$a\sum_{i=1}^{n}\boldsymbol{x}_i + bn = \sum_{i=1}^{n}\tilde{\mathbf{z}}_i. \tag{7}$$

Equations (6) and (7) can be put into matrix form, which is convenient for solutions to a and b via Least-Square or Maximum Likelihood Estimate (MLE) methods [1],

$$\begin{bmatrix} n & \sum_{i=1}^{n}\boldsymbol{x}_i \\ \sum_{i=1}^{n}\boldsymbol{x}_i & \sum_{i=1}^{n}\boldsymbol{x}_i^2 \end{bmatrix}\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n}\tilde{\mathbf{z}}_i \\ \sum_{i=1}^{n}\tilde{\mathbf{z}}_i\boldsymbol{x}_i \end{bmatrix}. \tag{8}$$

B. Maximum Likelihood

For measurements $\tilde{\mathbf{z}}_i$ for $i = 1\ldots n$, the measurement matrix $\mathbf{H}$ is

$$\mathbf{H} = \begin{bmatrix} 1 & \tilde{\mathbf{z}}_i \\ \vdots & \vdots \\ 1 & \tilde{\mathbf{z}}_n \end{bmatrix}, \tag{9}$$

the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta} = [b,\ a]^\intercal$ is

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^\intercal\mathbf{P}^{-1}\mathbf{H})^{-1}\mathbf{H}^\intercal\mathbf{P}^{-1}\tilde{\mathbf{z}}, \tag{10}$$

where $\mathbf{P} = \sigma^2\mathbf{I}$. The predicted measurement is

$$\hat{\mathbf{z}} = \mathbf{H}\hat{\boldsymbol{\theta}}. \tag{11}$$

The residual is

$$\mathbf{r} = \tilde{\mathbf{z}} - \hat{\mathbf{z}}, \tag{12}$$

with measurement covariance

$$\mathbf{C} = \mathbf{P} - \mathbf{H}^\intercal(\mathbf{H}^\intercal\mathbf{P}^{-1}\mathbf{H})^{-1}\mathbf{H}^\intercal. \tag{13}$$

C. R-Squared

The *Coefficient of Determination*, also known as R-Squared (r^2), is a common metric used to determine the variation of a data-fit. R-Squared values have a range $[0\cdots 1]$. If $r^2 = 1$ then all of the data-points lie perfectly on the regression line, and the predictor $\hat{\boldsymbol{\theta}}$ accounts for all of the variation in $\tilde{\mathbf{z}}$. If $r^2 = 0$, the predictor $\hat{\boldsymbol{\theta}}$ accounts for none of the variation in $\tilde{\mathbf{z}}$. This method requires that three values to be computed.

- SSE: Sum of squared errors (residuals), e.g. $\text{SSE} = \sum_{i=1}^{n}\mathbf{r}_i^2$.
- SSR: Regression sum of squares, the sum of squared deviations of the fitted values from their mean, e.g. $\text{SSR} = \sum_{i=1}^{n}(\hat{\mathbf{z}}_i - \bar{\mathbf{z}})^2$, where $\bar{\mathbf{z}}$ is the mean of $\sum_{i=1}^{n}\hat{\mathbf{z}}_i$.
- SST: Total sum of squares, the sum of squared deviations of y from mean(y), e.g. $\text{SST} = \text{SSE} + \text{SSR}$.

The R-squared value is computed by

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}. \tag{14}$$

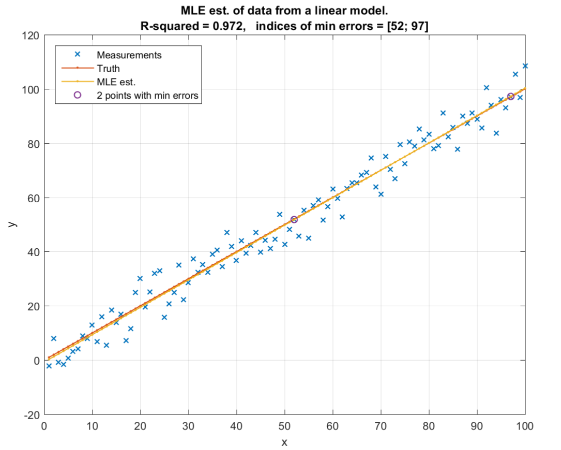

Fig. 1. MLE line-fit of noisy data, with 2 minimum error points identified.

D. *Minimum-risk data-fitting*

The objective of common outlier rejection algorithms [2] is to remove only measurements which are not consistent with a model or set of "valid" measurements. These methods by design utilize all measurements possible.

An alternative to this approach, is to instead utilize the minimum number of measurements required for a given model, thereby rejecting all measurements (including outliers) which are not required to fit the model.

A line-fit requires a minimum of two points, therefore the minimum-risk data-fit seeks the two points which best account for the variation in the measurements $\tilde{\mathbf{z}}$.

One possible method finds the two points with the minimum absolute value of the normalized residual (see eqn. (15)) by sorting the data in ascending order (see Section II of [**?**]) and selecting the two minimum values:

$$\mathbf{s} = \text{sort}\left(\left| \frac{\mathbf{r}}{\sqrt{\text{diag}(\mathbf{C})}} \right| \right). \tag{15}$$

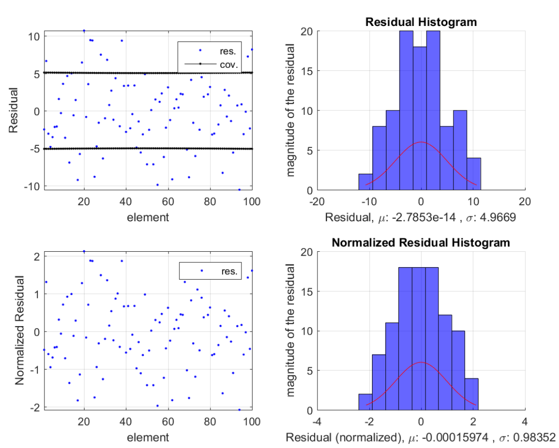

Fig. 2. MLE line-fit residuals and normalized residuals.

III. Illustrative Example

To illustrate the theory, a set of data was created for $\boldsymbol{x}$ and $\mathbf{y}$ using eqn. (1) for known a and b, measurement noise was added to produce $\tilde{\mathbf{z}}$, and the values for a and b were estimated using eqn. (11). The true line (eqn. (1)), measurements (eqn. (2)), MLE line-fit (eqn. (11)) and minimum error values (Section II-D) are shown in Fig. 1. The two minimum error values were obtained by sorting the absolute value of the residuals to find the two with the minimum error. The R-squared value was computed using eqn. (14). The MLE has an R-squared value of 0.972, indicating a good fit to the data.

References

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing, Vol. I - Estimation Theory*. Prentice Hall PTR, 2013.

[2] ——, *Fundamentals of Statistical Signal Processing, Vol. II - Detection Theory*. Prentice Hall PTR, 1998.