Training and Test Error of the OLS Estimator

DS-GA 1013 / MATH-GA 2824 Mathematical Tools for Data Science

Carlos Fernandez-Granda

Prerequisites

Mean-squared error estimation

Ordinary least squares (OLS)

OLS coefficient analysis

Quick recap

- Regression: Estimating response $\tilde{y}$ from features $\tilde{x}$
- Optimal estimator in mean squared error is conditional mean $\mathrm{E}(\tilde{y} \mid \tilde{x})$
- Unless features are very few, we can't compute it
- Linear models are interpretable and often very effective
- OLS estimator: $y_{\text{OLS}}(\tilde{x}) := \tilde{x}^T \beta_{\text{OLS}}$ (assuming everything is centered)

$$\beta_{\text{OLS}} := \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i^T \beta \right)^2$$

where (y_1, x_1), ..., (y_n, x_n) are training data

Quick recap

Analysis assuming data are indeed generated by linear model

$$\tilde{y} = \tilde{x}^T \beta_{\text{true}} + \tilde{z}$$

$\tilde{z}$ is Gaussian noise with standard deviation σ

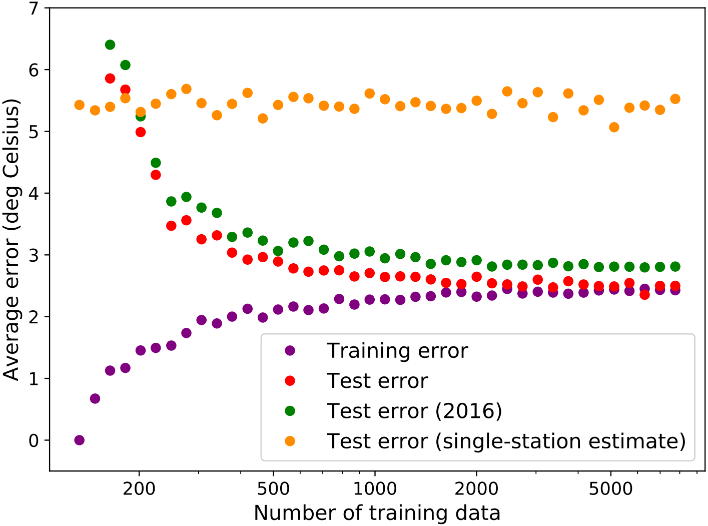
If we have access to joint distribution of $\tilde{x}$ and $\tilde{y}$, linear estimation achieves an error of σ

But we never have access to true distribution, only to samples (y_1, x_1), ..., (y_n, x_n)

Temperature prediction via linear regression

- Dataset of hourly temperatures measured at weather stations all over the US
- Goal: Predict temperature in Yosemite from other temperatures
- Response: Temperature in Yosemite
- Features: Temperatures in 133 other stations ($p = 133$) in 2015
- Test set: 10^3 measurements
- Additional test set: All measurements from 2016

Goal: Understand this



Model for training data

$$\tilde{y}_{\text{train}} := X^T \beta_{\text{true}} + \tilde{z}_{\text{train}}$$

- Feature matrix $X \in \mathbb{R}^{p \times n}$ is deterministic
- Coefficients $\beta_{\text{true}} \in \mathbb{R}^p$ are deterministic
- Noise $\tilde{z}_{\text{train}}$ is an n-dimensional iid Gaussian vector with zero mean and variance σ^2

OLS coefficient estimate

$$\beta_{\text{OLS}} = \beta_{\text{true}} + US^{-1}V^{T}\tilde{z}_{\text{train}}$$

Gaussian with mean β_{true} and covariance matrix $\sigma^2 US^{-2}U^T$

Error depends on singular values of feature matrix

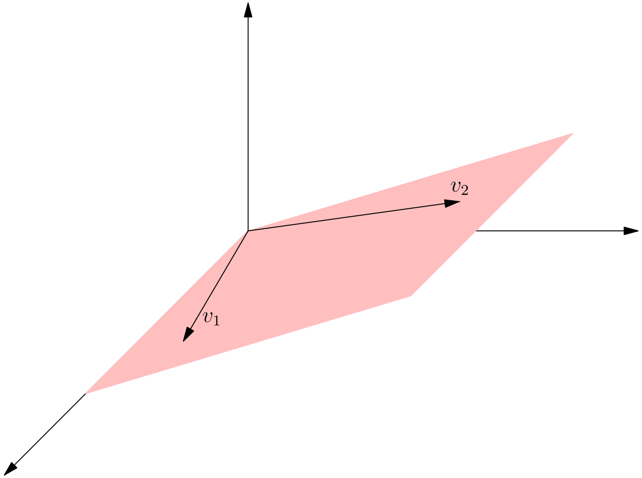
If singular values are small, error explodes!

What about the response?

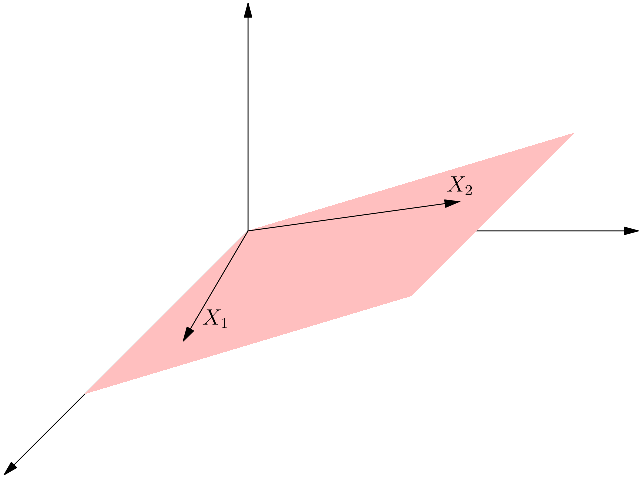
From a linear algebra perspective

$$
\begin{aligned}
X^T\beta &= \begin{bmatrix} x_1[1] & x_1[2] & \cdots & x_1[p] \\ x_2[1] & x_2[2] & \cdots & x_2[p] \\ \cdots & \cdots & \cdots & \cdots \\ x_n[1] & x_n[2] & \cdots & x_n[p] \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_p \end{bmatrix} \\
&= \beta_1 \begin{bmatrix} x_1[1] \\ x_2[1] \\ \cdots \\ x_n[1] \end{bmatrix} + \beta_2 \begin{bmatrix} x_1[2] \\ x_2[2] \\ \cdots \\ x_n[2] \end{bmatrix} + \cdots + \beta_p \begin{bmatrix} x_1[p] \\ x_2[p] \\ \cdots \\ x_n[p] \end{bmatrix} \\
&= \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p
\end{aligned}
$$

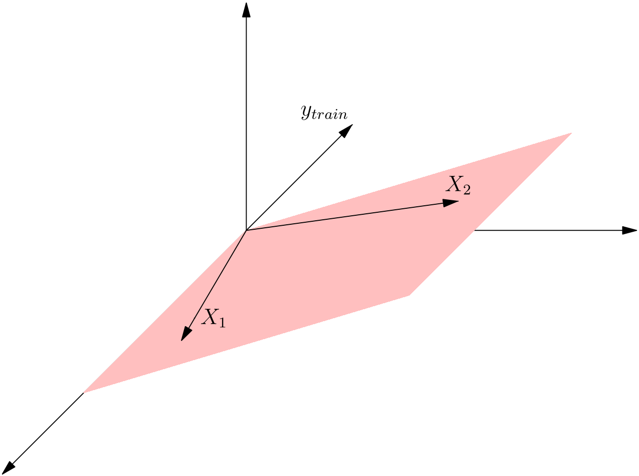
Subspace



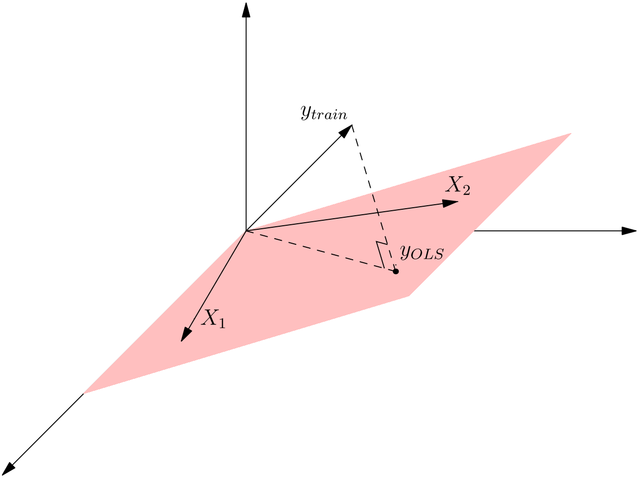
Linear model



Linear model



OLS estimate is a projection



Training error

$$\begin{aligned}
\tilde{y}_{\text{train}} - X^T \tilde{\beta}_{\text{OLS}} &= \tilde{y}_{\text{train}} - \mathcal{P}_{\text{row}(X)}\, \tilde{y}_{\text{train}} \\
&= X^T \beta_{\text{true}} + \tilde{z}_{\text{train}} - \mathcal{P}_{\text{row}(X)} \left(X^T \beta_{\text{true}} + \tilde{z}_{\text{train}} \right) \\
&= X^T \beta_{\text{true}} + \tilde{z}_{\text{train}} - X^T \beta_{\text{true}} - \mathcal{P}_{\text{row}(X)}\, \tilde{z}_{\text{train}} \\
&= \mathcal{P}_{\text{row}(X)^\perp}\, \tilde{z}_{\text{train}}
\end{aligned}$$

Goal: Characterize average training square error

$$\widetilde{E}^2_{\text{train}} := \frac{1}{n} \left|\left| \tilde{y}_{\text{train}} - X^T \tilde{\beta}_{\text{OLS}} \right|\right|_2^2$$
$$= \frac{1}{n} \left|\left| \mathcal{P}_{\text{row}(X)^\perp} \, \tilde{z}_{\text{train}} \right|\right|_2^2$$

Requires studying the projection of an iid Gaussian vector on a subspace

In $\mathbb{R}^n$ what fraction of variance is captured by subspace of dimension $n - p$? $\frac{n-p}{n}$

Average training square error

$$\left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\tilde{z}_{\text{train}}\right|\right|_2^2 = \left|\left|V_\perp V_\perp^T \tilde{z}_{\text{train}}\right|\right|_2^2$$
$$= \tilde{z}_{\text{train}}^T V_\perp V_\perp^T V_\perp V_\perp^T \tilde{z}_{\text{train}}$$
$$= \left|\left|V_\perp^T \tilde{z}_{\text{train}}\right|\right|_2^2$$

$V_\perp^T \tilde{z}_{\text{train}}$ is an $n-p$ dimensional Gaussian vector with covariance matrix

$$\Sigma_{V_\perp^T \tilde{z}_{\text{train}}} = V_\perp^T \Sigma_{\tilde{z}_{\text{train}}} V_\perp$$
$$= V_\perp^T \sigma^2 I V_\perp$$
$$= \sigma^2 I$$

It's an iid Gaussian vector!

ℓ_2 norm of d-dimensional iid standard Gaussian vector $\tilde{w}$

$$
\begin{aligned}
\mathrm{E}\left(||\tilde{w}||_2^2\right) &= \mathrm{E}\left(\sum_{i=1}^{d} \tilde{w}[i]^2\right) \\
&= \sum_{i=1}^{d} \mathrm{E}\left(\tilde{w}[i]^2\right) \\
&= d
\end{aligned}
$$

ℓ_2 norm of d-dimensional iid standard Gaussian vector $\tilde{w}$

$$\begin{aligned}
\mathrm{E}\left[\left(||\tilde{w}||_2^2\right)^2\right] &= \mathrm{E}\left[\left(\sum_{i=1}^{d} \tilde{w}[i]^2\right)^2\right] \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} \mathrm{E}\left(\tilde{w}[i]^2 \tilde{w}[j]^2\right) \\
&= \sum_{i=1}^{d} \mathrm{E}\left(\tilde{w}[i]^4\right) + 2\sum_{i=1}^{d-1}\sum_{j=i+1}^{d} \mathrm{E}\left(\tilde{w}[i]^2\right)\mathrm{E}\left(\tilde{w}[j]^2\right) \\
&= 3d + d(d-1) \quad \text{(4th moment of standard Gaussian} = 3) \\
&= d(d+2)
\end{aligned}$$

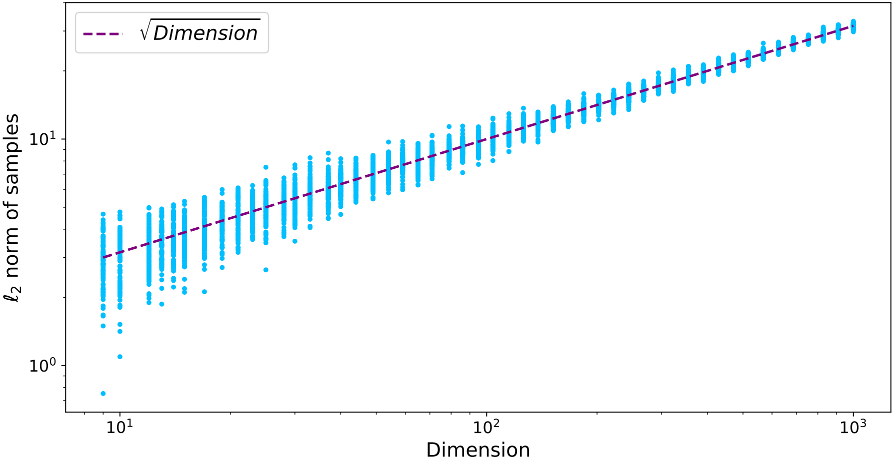
$$\begin{aligned}
\mathrm{Var}\left(||\tilde{w}||_2^2\right) &= \mathrm{E}\left[\left(||\tilde{w}||_2^2\right)^2\right] - \mathrm{E}^2\left(||\tilde{w}||_2^2\right) \\
&= 2d
\end{aligned}$$

ℓ_2 norm of d-dimensional iid standard Gaussian vector

As d grows, std / mean ratio of squared ℓ_2 norm scales as $1/\sqrt{d}$

Consequently squared ℓ_2 norm concentrates around d

ℓ_2 norm of d-dimensional iid standard Gaussian vector



Average training square error

$$\begin{aligned} \widetilde{E}^2_{\text{train}} &= \frac{1}{n} \left|\left| V_\perp^T \tilde{z}_{\text{train}} \right|\right|_2^2 \\ &= \frac{\sigma^2}{n} \left|\left| \tilde{w} \right|\right|_2^2 \end{aligned}$$

Dimension? $n - p$

$$\mathrm{E}\left(\widetilde{E}^2_{\text{train}}\right) = \sigma^2 \left(1 - \frac{p}{n}\right)$$

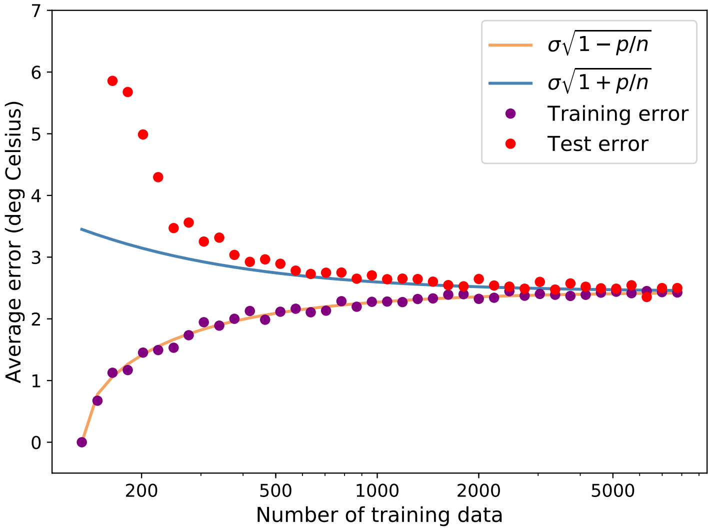
Average training square error

$$\text{Training error} \approx \sigma\sqrt{1 - \frac{p}{n}}$$

When $p << n$, error = noise

When $p \approx n$, error is very small: good news?

Observed training square error



Test data

Training data

$$\tilde{y}_{\text{train}} := X^T \beta_{\text{true}} + \tilde{z}_{\text{train}}$$

Test data

$$\tilde{y}_{\text{test}} := \tilde{x}_{\text{test}}^T \beta_{\text{true}} + \tilde{z}_{\text{test}}$$

$\tilde{x}_{\text{test}}$ is zero mean

$\tilde{z}_{\text{test}}$ is zero-mean Gaussian with variance σ^2

Test error

Goal: Characterize mean square of

$$\begin{aligned} \widetilde{E}_{\text{test}} &:= \tilde{y}_{\text{test}} - \tilde{x}_{\text{test}}^T \tilde{\beta}_{\text{OLS}} \\ &= \tilde{z}_{\text{test}} + \tilde{x}_{\text{test}}^T \left(\beta_{\text{true}} - \tilde{\beta}_{\text{OLS}} \right) \end{aligned}$$

where $\tilde{\beta}_{\text{OLS}}$ is computed from the training data

By independence

$$\mathrm{Var}\left(\tilde{y}_{\text{test}} - \tilde{x}_{\text{test}}^T \tilde{\beta}_{\text{OLS}} \right) = \sigma^2 + \mathrm{Var}\left(\tilde{x}_{\text{test}}^T \left(\beta_{\text{true}} - \tilde{\beta}_{\text{OLS}} \right) \right)$$

Everything is zero mean so mean square = variance

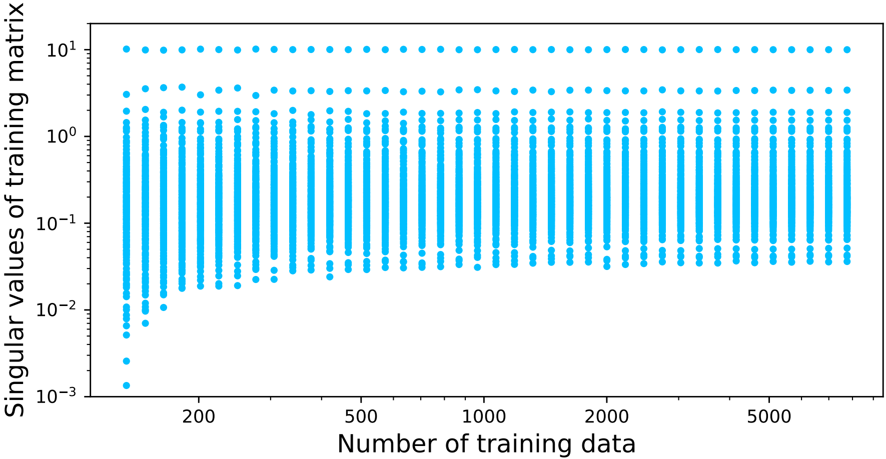
Coefficient error

Let USV^T be the SVD of X

$$\begin{aligned}
\beta_{\text{OLS}} - \beta_{\text{true}} &= (XX^T)^{-1}X\tilde{y} - \beta_{\text{true}} \\
&= (XX^T)^{-1}X(X^T\beta_{\text{true}} + \tilde{z}_{\text{train}}) - \beta_{\text{true}} \\
&= US^{-1}V^T\tilde{z}_{\text{train}} \\
&= \sum_{i=1}^{p} \frac{v_i^T\tilde{z}_{\text{train}}}{s_i} u_i
\end{aligned}$$

Potentially worrying: singular values can be very small

Singular values for temperature dataset



Mean square test error

$$\mathrm{E}\left[\left(\tilde{x}_{\text{test}}^T\left(\beta_{\text{true}}-\tilde{\beta}_{\text{OLS}}\right)\right)^2\right] = \mathrm{E}\left[\left(\sum_{i=1}^{p}\frac{v_i^T\tilde{z}_{\text{train}}\,u_i^T\tilde{x}_{\text{test}}}{s_i}\right)^2\right]$$

$$= \sum_{i=1}^{p}\frac{\mathrm{E}\left[(v_i^T\tilde{z}_{\text{train}})^2\right]\mathrm{E}\left[(u_i^T\tilde{x}_{\text{test}})^2\right]}{s_i^2}$$

$$\mathrm{E}\left(\frac{v_i^T\tilde{z}_{\text{train}}\,u_i^T\tilde{x}_{\text{test}}}{s_i}\,\frac{v_j^T\tilde{z}_{\text{train}}\,u_j^T\tilde{x}_{\text{test}}}{s_j}\right) = \frac{\mathrm{E}\left(u_i^T\tilde{x}_{\text{test}}u_j^T\tilde{x}_{\text{test}}\right)}{s_is_j}v_i^T\mathrm{E}\left(\tilde{z}_{\text{train}}\tilde{z}_{\text{train}}^T\right)v_j$$

$$= \frac{\mathrm{E}\left(u_i^T\tilde{x}_{\text{test}}u_j^T\tilde{x}_{\text{test}}\right)}{s_is_j}v_i^Tv_j$$

$$= 0 \qquad \text{for } i \neq j$$

Mean square test error

$$
\begin{aligned}
\mathrm{E}\left[\left(\tilde{x}_{\text{test}}^T\left(\beta_{\text{true}}-\tilde{\beta}_{\text{OLS}}\right)\right)^2\right] &= \sum_{i=1}^{p} \frac{\mathrm{E}\left[\left(v_i^T \tilde{z}_{\text{train}}\right)^2\right] \mathrm{E}\left[\left(u_i^T \tilde{x}_{\text{test}}\right)^2\right]}{s_i^2} \\
&= \sum_{i=1}^{p} \frac{v_i^T \mathrm{E}(\tilde{z}_{\text{train}} \tilde{z}_{\text{train}}^T) v_i u_i^T \mathrm{E}(\tilde{x}_{\text{test}} \tilde{x}_{\text{test}}^T) u_i}{s_i^2} \\
&= \sigma^2 \sum_{i=1}^{p} \frac{u_i^T \Sigma_{\tilde{x}_{\text{test}}} u_i}{s_i^2}
\end{aligned}
$$

$$
\mathrm{E}(\widetilde{E}_{\text{test}}^2) = \sigma^2 + \sigma^2 \sum_{i=1}^{p} \frac{\mathrm{Var}(u_i^T \tilde{x}_{\text{test}})}{s_i^2}
$$

Are small singular values problematic?

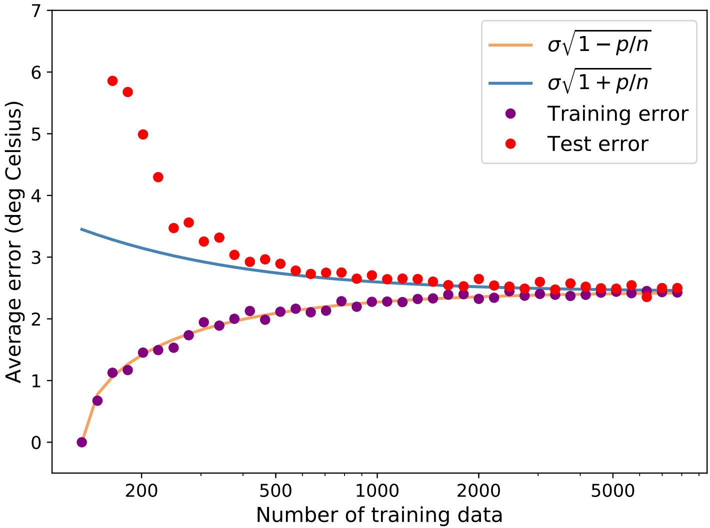
Mean square test error

$$\begin{aligned}\frac{s_i^2}{n} &= \frac{u_i U S^2 U^T u_i}{n} \\ &= \frac{u_i X X^T u_i}{n} \\ &= u_i^T \Sigma_{\mathcal{X}} u_i \\ &= \text{var}\left(\mathcal{P}_{u_i}\,\mathcal{X}\right)\end{aligned}$$

$$\begin{aligned}\mathrm{E}(\widetilde{E}_{\text{test}}^2) &= \sigma^2 + \sigma^2 \sum_{i=1}^{p} \frac{\mathrm{Var}(u_i^T \tilde{x}_{\text{test}})}{s_i^2} \\ &\approx \sigma^2 \left(1 + \frac{p}{n}\right)\end{aligned}$$

If variance estimated from training data $\approx$ test variance, small singular values are not a problem!

Observed test mean square error



What have we learned?

- Fitting a linear regression model can be interpreted in terms of a projection onto a subspace
- This yields a precise description of the training error as a function of the number of data
- If data are not enough we overfit!
- Test error can be low even if coefficient error is high, as long as data are enough to accurately estimate the covariance matrix of the features