Weighted Least Squares Problems

Statistical interpretation

Niclas Börlin

5DA001 Non-linear Optimization

Weighted least squares

Formulation

- In the least squares problem

$$\min_x \frac{1}{2}\|r(x)\|_2^2 = \min_x \frac{1}{2} r(x)^T r(x),$$

  each element of $r(x)$ carry the same weight.
- A more general formulation is

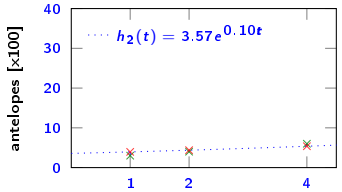
$$\min_x \frac{1}{2}\|r(x)\|_W^2 = \min_x \frac{1}{2} r(x)^T W r(x),$$

  where the weight matrix W is symmetric positive semi-definite.
- *How do we choose the weights?*

Statistical interpretation

Stochastic model

- Consider our antelope problem:



- If the residuals are interpreted statistically, i.e. we have a model

$$y_i = x_1 e^{x_2 t_i} + \varepsilon_i,$$
$$r_i = x_1 e^{x_2 t_i} - y_i,$$

  we may make statistical assumptions about the errors ε_i and residuals r_i.
- A common statistical model is that the errors are normally distributed and optionally indepedent.

Weighted least squares

Choice of weights, independent observations

- Good observation, i.e. with small uncertainty, should be given a larger weights than bad ones.
- If each residual $r_i(x)$ is independent and $N(0, \sigma_i^2)$, the weights

$$w_{ii} = \frac{1}{\sigma_i^2}, \qquad w_{ij} = 0, i \neq j,$$

  are optimal and will give the maximum likelyhood estimators of x given our observations.
- The objective function becomes

$$\begin{aligned}\frac{1}{2} r(x)^T W r(x) &= \frac{1}{2}\left(w_1 r_1(x)^2 + \cdots + w_m r_m(x)^2\right) \\ &= \frac{1}{2}\left(\frac{r_1(x)^2}{\sigma_1^2} + \cdots + \frac{r_m(x)^2}{\sigma_m^2}\right).\end{aligned}$$

- Thus, observations with small uncertainties σ_i are given large weights $1/\sigma_i$, and vice versa.

Weighted least squares

The general case

- If the residual vector is assumed to be

$$r \sim N(0, C),$$

where the symmetric positive semi-definite matrix C is the *covariance matrix* for r, the optimal weight matrix is

$$W = C^{-1}.$$

- The independent case corresponds to diagonal C and W.
- The distance measure $r(x)^T C^{-1} r(x)$ is sometimes called the Mahalanobis distance.

Weighted least squares

Methods

- If we want to solve a weighted least squares problem, there are two equivalent solutions:
  - Modify the algorithm.
  - Modify the residual/Jacobian function.
- A modified algorithm would solve the following equation

$$J^T W J p = -J^T W r$$

at every iteration.

- A modified residual/Jacobian would be

$$r_s(x) = L^T r(x),$$
$$J_s(x) = L^T J(x),$$

where $LL^T = W$ is the Cholesky factorization of W.

- Such a factor L will always exist if W is positive semidefinite.

Linear error propagation

- Given a random variable

$$\underline{\mathbf{x}} \sim N(\mu_x, \mathbf{C}_{xx}),$$

and a linear transformation

$$\underline{\mathbf{y}} = \mathbf{A}\underline{\mathbf{x}} + \mathbf{b},$$

the transformed variable is

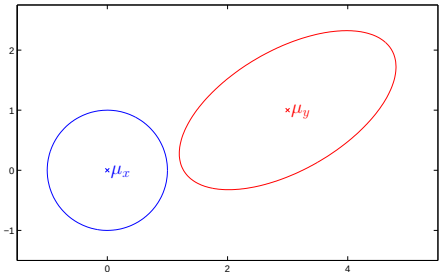
$$\underline{\mathbf{y}} \sim N(\mathbf{A}\mu_x + \mathbf{b}, \mathbf{A}\mathbf{C}_{xx}\mathbf{A}^T),$$

i.e.

$$\mu_y = \mathbf{A}\mu_x + \mathbf{b},$$
$$\mathbf{C}_{yy} = \mathbf{A}\mathbf{C}_{xx}\mathbf{A}^T.$$

Example (Scale, rotate, and shift)

$$\underline{\mathbf{x}} \sim N(\mathbf{0}, \mathbf{I}_2), \qquad \underline{\mathbf{y}} = \mathbf{A}\underline{\mathbf{x}} + \mathbf{b},$$
$$\mathbf{A} = \begin{pmatrix} \cos 30^\circ & -\sin 30^\circ \\ \sin 30^\circ & \cos 30^\circ \end{pmatrix} \begin{pmatrix} 2 & \\ & 1 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}.$$




```
>> x=zeros(2,1); Cxx=eye(2); y=A*x+b; Cyy=A*Cxx*A';
>> [V,D]=eig(Cyy); S=sqrt(D); t=linspace(0,2*pi); cx=[cos(t);sin(t)];
>> px=repmat(x,1,length(t))+I*cx; py=repmat(y,1,length(t))+V'*S*cx;
>> plot(px(1,:),px(2,:),'b',py(1,:),py(2,:),'r'), axis equal
```

Covariance and correlation

- The covariance σ_{xy} describe the co-variation between the errors in x and y.
- It is difficult to determine if a covariance value σ_{xy} is large or not, since it depend on the size of σ_x and σ_y.
- However, if the covariance is normalized by the standard deviations, we get the *correlation coefficient* ρ_{xy}, defined as

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \in [-1, +1].$$

- If $\rho_{xy} = 0$, the variables $\underline{x}$ and $\underline{y}$ are said to be *uncorrelated*.
- High $|\rho_{xy}|$ values imply that the variables $\underline{x}$ and $\underline{y}$ are (almost) linearly dependent, i.e. they cannot be estimated independently.

Computation of correlation coefficients

- If $\mathbf{D}$ is the diagonal part of the covariance matrix $\mathbf{C}$ and

$$\mathbf{S} = \mathbf{D}^{-1/2} = \begin{pmatrix} \frac{1}{\sqrt{c_{11}}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{c_{nn}}} \end{pmatrix},$$

the matrix of correlation coefficients $\mathbf{P}$ may be computed as

$$\mathbf{P} = \mathbf{S}\mathbf{C}\mathbf{S}^T.$$

Example

```
>> Cxx, S=diag(1./sqrt(diag(Cxx))); RHO=S*Cxx*S'
Cxx =
     1     0
     0     1
RHO =
     1     0
     0     1
>> Cyy, S=diag(1./sqrt(diag(Cyy))); RHO=S*Cyy*S'
Cyy =
    3.2500    1.2990
    1.2990    1.7500
RHO =
    1.0000    0.5447
    0.5447    1.0000
```


Non-linear error propagation

- Given a non-linear function $\underline{\mathbf{y}} = \mathbf{g}(\underline{\mathbf{x}})$ and its Taylor expansion

$$\underline{\mathbf{y}} = \mu_{\mathbf{y}} + d\underline{\mathbf{y}} = \mathbf{g}(\mu_{\mathbf{x}}) + \mathbf{J}d\underline{\mathbf{x}} + \mathcal{O}(\|d\underline{\mathbf{x}}\|^2),$$

where the Jacobian is

$$\mathbf{J} = [J_{ij}] = \left[\frac{\partial g_i(\mathbf{x})}{\partial x_j}\right]_{\underline{\mathbf{x}}=\mu_{\mathbf{x}}},$$

we get a *first order approximation* of the distribution of $\underline{\mathbf{y}}$ as

$$\mu_{\mathbf{y}} = \mathbf{g}(\mu_{\mathbf{x}}),$$
$$\mathbf{C}_{yy} = \mathbf{J}\mathbf{C}_{xx}\mathbf{J}^T.$$

Non-linear error propagation

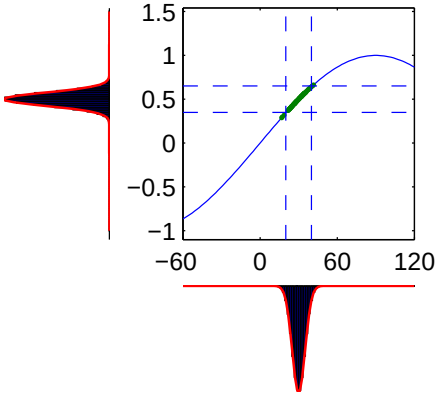
Good approximation

- If the function is approximately linear in $\mu_x \pm k\sigma_x$, the approximation is good:
- For

$$\underline{y} = \sin \underline{x},$$
$$\mu_x = \pi/6 \ (30^\circ),$$
$$\sigma_x = \pi/36 \ (5^\circ),$$

we get

$$\mu_y = \sin \pi/6 = 0.5,$$
$$\mathbf{J} = \cos \pi/6,$$
$$\sigma_y = \sqrt{\mathbf{J}\sigma_x^2\mathbf{J}^T} \approx 0.08.$$

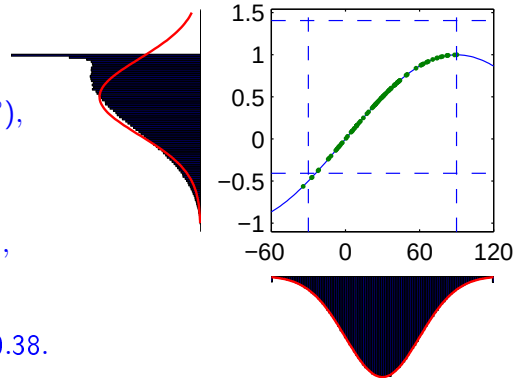


Non-linear error propagation

Poor approximation

- Otherwise, the approximation is poor:
- For

$$\underline{y} = \sin \underline{x},$$
$$\mu_x = \pi/6 \ (30^\circ),$$
$$\sigma_x = 5\pi/36 \ (25^\circ),$$



we get

$$\mu_y = \sin \pi/6 = 0.5,$$
$$\mathbf{J} = \cos \pi/6,$$
$$\sigma_y = \sqrt{\mathbf{J}\sigma_x^2\mathbf{J}^T} \approx 0.38.$$

Example (Polar-to-cartesian conversion)

- A vector with polar coordinates

$$\mathbf{z} = \begin{pmatrix} \theta \\ r \end{pmatrix}$$

has corresponding cartesian coordinates

$$\mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r\cos\theta \\ r\sin\theta \end{pmatrix}.$$

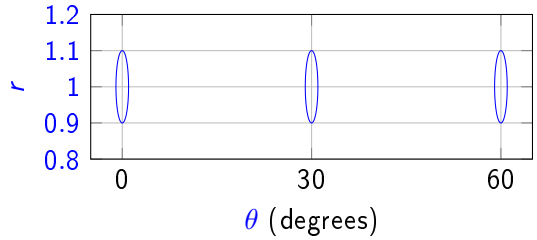
- The transformation has Jacobian

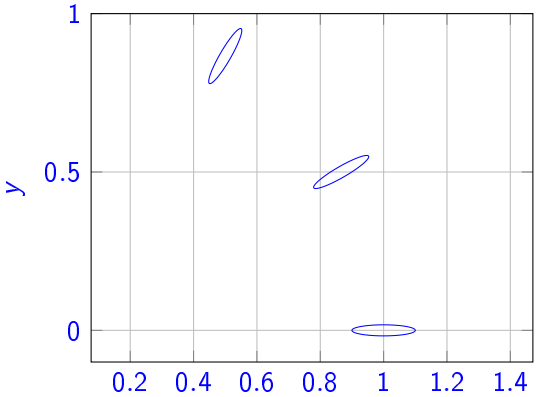
$$\mathbf{J} = \begin{pmatrix} -r\sin\theta & \cos\theta \\ r\cos\theta & \sin\theta \end{pmatrix}$$

and

$$\mathbf{C}_{vv} = \mathbf{J}\mathbf{C}_{zz}\mathbf{J}^T.$$

Example (Polar-to-cartesian conversion)





Linear estimation

Model

- Assume we have a vector $\mathbf{b}$ that is an observation of a stochastic vector

$$\underline{\mathbf{b}} \sim N(\mu_b, \mathbf{C}_{bb}).$$

- Furthermore, assume that the "exact" observation vector μ_b is explained by a linear model

$$\mathbf{A}\mu_x = \mu_b,$$

for some unknown value of the parameter vector μ_x.
- The $m \times n$-matrix *design matrix* $\mathbf{A}$, $m \geq n$, is assumed to be of full rank.
- The residual difference between the observations and what can be explained by the model is thus given by

$$\underline{\mathbf{v}} = \underline{\mathbf{b}} - \mu_b = \underline{\mathbf{b}} - \mathbf{A}\mu_x \sim N(\mathbf{0}, \mathbf{C}_{bb}).$$

Linear estimation

Optimal estimate

- If we choose to minimize the normalized residuals

$$\Omega^2 = \mathbf{v}^T \mathbf{C}_{bb}^{-1} \mathbf{v} = \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_{\mathbf{C}_{bb}^{-1}} = (\mathbf{b} - \mathbf{A}\mathbf{x})^T \mathbf{C}_{bb}^{-1} (\mathbf{b} - \mathbf{A}\mathbf{x}),$$

  we end up with the *weighted normal equations*

$$\mathbf{A}^T \mathbf{W} \mathbf{A}\mathbf{x} = \mathbf{A}^T \mathbf{W} \mathbf{b},$$

  with

$$\mathbf{W} = \mathbf{C}_{bb}^{-1}$$

  as the *weight matrix*.
- The estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ from the weighted normal equations is mathematically

$$\hat{\mathbf{x}} = \underbrace{(\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W}}_{=\mathbf{P}} \mathbf{b} = \mathbf{P}\mathbf{b}.$$

Linear estimation

Covariance of estimate

- If the design matrix $\mathbf{A}$ is exact, the covariance becomes

$$\mathbf{C}_{\hat{x}\hat{x}} = \mathbf{P}\mathbf{C}_{bb}\mathbf{P}^T = (\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{W}\mathbf{C}_{bb}\mathbf{W}\mathbf{A}(\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1}.$$

- If $\mathbf{W} = \mathbf{C}_{bb}^{-1}$ the covariance $\mathbf{C}_{\hat{x}\hat{x}}$ can be simplified to

$$\begin{aligned}\mathbf{C}_{\hat{x}\hat{x}} &= (\mathbf{A}^T\mathbf{C}_{bb}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{C}_{bb}^{-1}\mathbf{C}_{bb}\mathbf{C}_{bb}^{-1}\mathbf{A}(\mathbf{A}^T\mathbf{C}_{bb}^{-1}\mathbf{A})^{-1} \\ &= (\mathbf{A}^T\mathbf{C}_{bb}^{-1}\mathbf{A})^{-1}.\end{aligned}$$

- In this case, the $\hat{\mathbf{x}}$ estimate is also the *Maximum-likelyhood estimate* of $\mathbf{x}$.

Linear estimation

Relative weights

- A common case is when the true covariance $\mathbf{C}_{bb}$ is known *up to a scale factor*, i.e. $\mathbf{C}_{bb}$ may be written as

$$\mathbf{C}_{bb} = \sigma_0^2 \mathbf{S}_{bb},$$

  where the structure matrix $\mathbf{S}_{bb}$ is known but the *variance factor* σ_0^2 is not.
- Estimating $\hat{\mathbf{x}}$ with $\mathbf{W} = \mathbf{S}_{bb}^{-1}$ still yields the Maximum- likelyhood estimate, since

$$\begin{aligned}\hat{\mathbf{x}} &= (\mathbf{A}^T\mathbf{S}_{bb}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_{bb}^{-1}\mathbf{b} = \frac{\sigma_0^2}{\sigma_0^2}(\mathbf{A}^T\mathbf{S}_{bb}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_{bb}^{-1}\mathbf{b} \\ &= (\mathbf{A}^T\mathbf{C}_{bb}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{C}_{bb}^{-1}\mathbf{b}.\end{aligned}$$

- Thus, the optimal estimate only requires that the relative size of the errors is known.

Linear estimation

Estimating the variance factor

- In order to estimate the covariance $\mathbf{C}_{\hat{x}\hat{x}}$ of the estimate

$$\mathbf{C}_{\hat{x}\hat{x}} = (\mathbf{A}^T\mathbf{C}_{bb}^{-1}\mathbf{A})^{-1} = \sigma_0^2(\mathbf{A}^T\mathbf{S}_{bb}^{-1}\mathbf{A})^{-1},$$

  we need an estimate of σ_0^2.
- Given $\hat{\mathbf{x}}$ and

$$\hat{\mathbf{v}} = \mathbf{b} - \mathbf{A}\hat{\mathbf{x}},$$

  the variance factor σ_0^2 may be estimated as

$$\hat{\sigma}_0^2 = \frac{\hat{\mathbf{v}}^T\mathbf{S}_{bb}^{-1}\hat{\mathbf{v}}}{r},$$

  where the *redundancy* r is

$$r = m - n.$$

- This enables us to estimate $\mathbf{C}_{\hat{x}\hat{x}}$ as

$$\widehat{\mathbf{C}_{\hat{x}\hat{x}}} = \hat{\sigma}_0^2(\mathbf{A}^T\mathbf{S}_{bb}^{-1}\mathbf{A})^{-1}.$$

Linear estimation

The standard deviation of unit weight

- Two common cases:
  1. The covariance $\mathbf{C}_{bb} = \mathbf{S}_{bb}$ is known fairly well.
     - In this case $\sigma_0^2 = 1$ and the estimated value $\hat{\sigma}_0$ can be used to test our assumption.
     - The constant σ_0 is known as the *standard deviation of unit weight*.
  2. We assume the observations are independent with unknown variance, i.e. $\mathbf{S}_{bb} = \mathbf{I}$.
     - In this case, $\hat{\sigma}_0$ will be an estimate of the measurement error for each element.

Linear estimation

Contribution

- Consider the estimated observations

$$\hat{\mathbf{b}} = \mathbf{A}\hat{\mathbf{x}} = \underbrace{\mathbf{AP}}_{=\mathbf{U}}\mathbf{b} = \mathbf{Ub}.$$

- The contribution matrix $\mathbf{U}$ contain elements $0 \leq u_{ij} \leq 1$ that describe how much of observations j contribute to the estimated value i.

Linear estimation

Redundancy

- Consider the estimated residuals

$$\hat{\mathbf{v}} = \mathbf{b} - \mathbf{A}\hat{\mathbf{x}} = \mathbf{b} - \mathbf{Ub} = \underbrace{(\mathbf{I} - \mathbf{U})}_{=\mathbf{R}}\mathbf{b} = \mathbf{Rb}.$$

- The redundancy matrix $\mathbf{R}$ contain elements $0 \leq r_{ij} \leq 1$ that describe how much of observation j affects residual i.
- The elements r_{ii} is the *redundancy number* for observation i and

$$\operatorname{tr}\mathbf{R} = \operatorname{rk}\mathbf{R} = r = m - n.$$

- A high redundancy number means that a large error in observation i will generate a large residual v_i. This simplifies blunder detection.
- A low redundancy number means that a large error in observation i will be invisible in the residual v_i, making blunder detection difficult.
- Points with low redundandy numbers are called leverage points.

Statistical interpretation

Variance of estimated parameters

- The variance for the estimated parameters are calculated from the variance-covariance matrix

$$D = \sigma^2(\nabla^2 f(x^*))^{-1},$$

where each diagonal element d_{ii} correspond to the variance of the parameter x_i, and the off-diagonal element d_{ij} correspond to the covariance between parameters x_i and x_j.

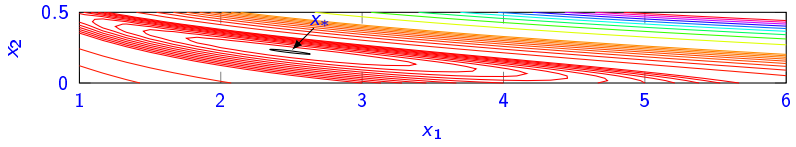
Statistical interpretation

Variance of estimated parameters

- A high variance means a high degree of uncertainty about a parameter.
- In this context, the inverse matrix

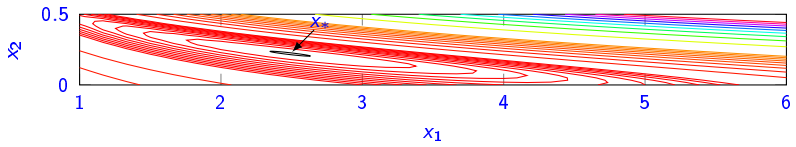
$$K = D^{-1} = \frac{1}{\sigma^2}\nabla^2 f(x^*),$$

  is sometimes called the information matrix, since higher diagonal values k_{ii} correspond to more information about the parameter x_i.
- Since the information matrix is proportional to the hessian $H(x^*) = \nabla^2 f(x^*)$, strong curvature corresponds to high information content, i.e. good localization of the parameter.



Statistical interpretation

Variance of estimated parameters



$$x^* = \begin{bmatrix} 2.49 \\ 0.22 \end{bmatrix}, r(x^*) = \begin{bmatrix} 0.11 \\ -0.13 \\ 0.03 \end{bmatrix}, J(x^*) = \begin{bmatrix} 1.25 & 3.11 \\ 1.56 & 7.75 \\ 2.42 & 24.1 \end{bmatrix}, J(x^*)^T J(x^*) = \begin{bmatrix} 9.84 & 74.3 \\ 74.3 & 651 \end{bmatrix},$$

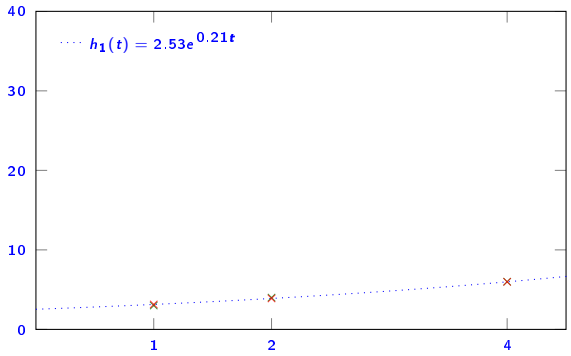
$$Q(x^*) = \begin{bmatrix} 0 & -10^{-3} \\ -10^{-3} & 0.96 \end{bmatrix}, H(x^*) = J(x^*)^T J(x^*) + Q(x^*) = \begin{bmatrix} 9.84 & 74.3 \\ 74.3 & 652 \end{bmatrix},$$

$$H(x^*)^{-1} = \begin{bmatrix} 0.73 & -0.083 \\ -0.083 & 0.011 \end{bmatrix} = V\Lambda V^T, V = \begin{bmatrix} 0.99 & 0.11 \\ -0.11 & 0.99 \end{bmatrix}, \Lambda = \begin{bmatrix} 0.74 & 0 \\ 0 & 0.0015 \end{bmatrix},$$

- Thus, $\hat{\sigma} = \sqrt{r(x^*)^T r(x^*)/(3-2)} = 0.17$ (hecto-antelopes) and the standard deviation of x_1 is $\sqrt{0.73}\sigma = 0.14$ (hecto-antelopes) and of x_2 is $\sqrt{0.011}\sigma = 0.017$ (hecto-antelopes/year). With these units, the maximum uncertainty is in the direction of $0.99x_1 - 0.11x_2$.
- Note that the interpretation of the standard deviations is context-dependent, since it depends on e.g. the measurement units of each parameter.

Statistical interpretation

Redundancy numbers



$$\mathbf{U} = \begin{pmatrix} 0.60 & 0.48 & -0.10 \\ 0.48 & 0.42 & 0.12 \\ -0.10 & 0.12 & 0.97 \end{pmatrix}, \mathbf{R} = \begin{pmatrix} \mathbf{0.40} & -0.48 & 0.10 \\ -0.48 & \mathbf{0.58} & -0.12 \\ 0.10 & -0.12 & \mathbf{0.03} \end{pmatrix}.$$