When we formulate and solve a nonlinear least squares problem, we need to use a numerical optimization routing to find the maximum likelihood estimate. In order to do this, we first scale the vector of model-measurement mismath residuals $y = M(\theta)$ by using a - usually diagonal - guess Σ_{s} of the covariance matrix of the noise, it

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM

order to obtain the scaled residual vector $R(\theta) := \Sigma_{n}^{-\frac{1}{2}} (M(\theta) - y)$ such that the maximum likelihood estimate $\hat{\theta} = \theta^{-}$ is obtained by the solution of the optimization nucleum.

 $\theta^* = \arg \min_{\theta} \frac{\frac{1}{2} \|R(\theta)\|_2^2}{\|R(\theta)\|_2^2}$

(5.30)

(5.31)

(5.32)

Note that the residual function R maps from \mathbb{R}^d to \mathbb{R}^N , and that most solution algorithms require the user to pass this function $R(\theta) =$ and not the objective function $f(\theta) = \frac{1}{4} \|R(\theta)\|_2^2$ – to the solver. We want to answer three questions in this section:

How do we solve the nonlinear least squares optimization problem (5.30) in practice?
 How can we obtain an estimate of the covariance matrix of the parameter estimate?

How can we assess if the modelling assumptions, in particular on the noise, were correct?

We will answer the three questions in the following three subsections.

5.5.1 The Gauss-Newton Algorithm

The second obtained on the spectrum sets on the spectra of the sp

Here: Because we know very well how to selve linear least squares problem and because all emiliance functions can locally be approximately their first order Tayles series, a simplificient di fast sould be to solve a linear least squares problem based on the linearization at a solution guess θ_{\pm} in order to obtain a better solution gues θ_{\pm} . More concretely, for any solution guess θ_{\pm} we have that $R(\theta) = R(\theta_{\pm}) + \frac{R}{2}(\theta_{\pm})(\theta - \theta_{\pm}) + O((\theta - \theta_{\pm})))$ and if we use the first order Taylor series to formulate a linear least squares problem in order to find $\theta_{\pm,0,\pm}$ by which the concretely, for a solution guess θ_{\pm} in order to substance the solution of the solution of

 $\begin{array}{ll} \theta_{|k+1|} &=& \arg\min_{\theta} \frac{1}{2} \left\| R(\theta_{|k|}) + \frac{\partial R}{\partial \theta} (\theta_{|k|}) (\theta - \theta_{|k|}) \right\|_{2}^{2} \\ &=& \arg\min_{\theta} \frac{1}{2} \left\| -J(\theta_{|k|}) \theta_{|k|} + R(\theta_{|k|}) + J(\theta_{|k|}) \theta \right\|_{2}^{2} \end{array}$

 $= (J(\theta_{kj})^\top J(\theta_{kl}))^{-1} J(\theta_{kl})^{-1} J(\theta_{kl}) \theta_{kl} - \frac{-\theta_{kl}}{\theta_{kl}}) (5.33)$ $= \theta_{kl} - (J(\theta_{kl})^\top J(\theta_{kl}))^{-1} J(\theta_{kl})^{-1} \theta_{kl}) = (5.34)$ $= \theta_{kl} - J(\theta_{kl})^{-1} R(\theta_{kl}) = (5.35)$ hore is only well defined if the Larobhan matrix $J(\theta_{kl})$ is of full rack, but that in trans-

Note that there territories anovel is volvely were advanced to according to an order $J(m_0)$ to see that the M_0 is a set of the territories o

52	CHAPTER 5. MAXIMUM LIKELIHOOD AND BAYESI	AN ESTIMATION
f, which are given by		
	$f(\theta) = \frac{1}{2} R(\theta) _2^2 = \frac{1}{2} \sum_{i=1}^{N} R_i(\theta)^2$	(5.36)
	$\nabla f(\theta) = \sum_{i=1}^{N} \nabla R_i(\theta) R_i(\theta) = J(\theta)^{\top} R(\theta)$	(5.37)

 $\nabla f(\theta) = \sum_{i=1}^{n} \nabla R_i(\theta) R_i(\theta) = J(\theta)^\top R(\theta)$ (5.37) $\nabla^2 f(\theta) = \underbrace{J(\theta)^\top J(\theta)}_{-\infty^{(n-1)}} + \sum_{i=1}^{N} \nabla^2 R_i(\theta) R_i(\theta)$ (5.38)

s, the iterations of the Gauss-Newton algorithm can be w

 $\boldsymbol{\theta}_{[k+1]} = \boldsymbol{\theta}_{[k]} - \boldsymbol{B}_{\mathrm{GN}}(\boldsymbol{\theta}_{[k]})^{-1} \nabla f(\boldsymbol{\theta}_{[k]})$

It can be seen, as expected from a optimization algorithm, that the algorithm would not move says from substrategy constraints $P(\theta_{0,0}) = 0$. But of the strategy constraints $P(\theta_{0,0}) = 0$. But of the strategy constraints of the strategy constrategy constraints of the strategy constrategy constrateg

5.5.2 Estimating the Covariance Matrix and Extracting its Relevant Entries

The easier way to obtain a rough estimate of the covariance matrix Σ_0 of the parameter estimate ℓ^* would be to summe that the lanceation of ℓ at at host other is the coverten model, and that all the statical assumptions we made in the formulation of the fraction R were coverte. i.e., that we indeed hold Gaussian noise with covariance matrix Σ_0 . Following the linear last systems analysis, we could then directly use $R(R)^{(1)} = \mu_0$ can summption where $L_0 = R(R) =$

 $\Sigma_{ij} := \frac{\|R(\theta^*)\|_2^2}{N-d} \left(J(\theta^*)^- J(\theta^*)\right)^{-1}$ is wants to express the result of the parameter estimation, one often only uses the diagonal entries from this is, which contain, for $i = 1, ..., d_i$ he variances σ_i^2 of the respective parameter components θ_{ij} as seen in the

ession: $\Sigma_{\vec{\theta}} = \begin{bmatrix} \sigma_1^2 & * & * & * \\ * & \sigma_2^2 & * & * \\ * & * & \ddots & * \\ * & * & * & \sigma_d^2 \end{bmatrix}$

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM

Taking the square root of the variances yields the standard deviations, such that the final result of the whole parameter estimation procedure could be presented by only 2d numbers in the form $\theta_i = \theta_i^* \pm \sqrt{\sigma_i^2}$, for i = 1, ..., d

5.5.3 Checking the Optimal Residual Vector

The second prior of the second prior of the second second prior of the second second prior of the second se

is to look at the residual values $R_i(\theta^*)$ for i = 1, ..., N. If we plot them as a function of i, they should look file a sequence of random sumbers. Is order to check this is more detail, one orgenized returns and pitors a histogram of the residual values $R_i(\theta^*)$. If the histogram look like the histogram of a zero mean Gaussian with invariance, the model assumptions on the system and noise are likely to be cornect. If and, some gut of the modelling assumptions was probably werego. One should then think head and change the system or tokes model and neurat the gammater estimation processing. In some of the system of the system of the set of the system of th

MSI Lecture 5.5 *Practical* solution of the Nonlinear Least Squares Problem

Per Rutquist

When we formulate and solve a nonlinear least squares problem, we need to use a numerical optimization routi to find the maximum likelihood estimate. In order to do this, we first scale the vector of model-measureme insmath residuals $y = M(\theta)$ by using a - usually diagonal - guess Σ_c of the covariance marks of the noise,

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM

order to obtain the scaled residual vector $R(\theta) := \Sigma_{\epsilon}^{-\frac{1}{2}} (M(\theta) - y)$ such that the maximum likelihood

 $\theta^* = \arg \min_{\theta} \frac{1}{2} ||R(\theta)||_2^2$

Note that the residual function R maps from \mathbb{R}^d so \mathbb{R}^N , and that most solution algorithms require the user to p this function $R(\theta)$ – and not the objective function $f(\theta) = \frac{1}{2} ||R(\theta)||_2^2$ – to the solver. We want to answer the constitution in this section:

How do we solve the nonlinear least squares optimization problem (5.30) in practice?
 How can we obtain an estimate of the covariance matrix of the measurement estimate?

How can we assess if the modelling assumptions, in particular on the noise, were correct

We will answer the three questions in the following three subsections.

5.5.1 The Gauss-Newton Algorithm

Idea: Because we know very well how to selve linear least squares problems and because all nonlinear function calculable paraproximately the first for each rate in a simplification of itera would be to solve a linear line squares problem haved on the intenzization at a solution guess $\theta_{\pm,1}$ in order to orbain a better solution guess $\theta_{\pm,1}$ we deter to orbain a better solution guess $\theta_{\pm,1}$ we deter to orbain a better solution guess $\theta_{\pm,1}$ in order to orbain a better solution guess $\theta_{\pm,1}$ and at $R(\theta) = R(\theta_{\pm,1}) = \frac{2\pi}{2\theta}(\theta_{\pm,1})(\theta - \theta_{\pm,1}) = O((\theta - \theta_{\pm,1}))$.

 $\theta_{[k+1]} = \arg \min_{\theta} \frac{1}{2} \left\| R(\theta_{[k]}) + \frac{\partial R}{\partial \theta}(\theta_{[k]})(\theta - \theta_{[k]}) \right\|_{2}^{2}$ = $\arg \min_{\theta} \frac{1}{2} \left\| -J(\theta_{[k]}) \theta_{[k]} + R(\theta_{[k]}) + J(\theta_{[k]}) \theta \right\|_{2}^{2}$

 $= (J(\theta_{[k]})^{-1}J(\theta_{[k]}))^{-1}J(\theta_{[k]})^{-1}(J(\theta_{[k]}), \theta_{[k]} - R(\theta_{[k]})) \qquad (5.33)$ $= \theta_{[k]} - (J(\theta_{[k]})^{-1}J(\theta_{[k]}))^{-1}J(\theta_{[k]})^{-1}R(\theta_{[k]}) \qquad (5.34)$ $= \theta_{[k]} - J(\theta_{[k]})^{-1}R(\theta_{[k]}) \qquad (5.35)$

tical informations, sumil algorithm modifications ensure that each interaction is well defined. With the above expression, we have simily defined the basis Gauss-Neutron indignition. One can show that -i if is converges – Gauss-Neutron algorithm converges linearly to a stationary point θ^* with $\nabla f(\theta^*) = 0$, but a proof of this result is byouth our interval. The shows the term However, is order to understand the algorithm a bit better and to see at least why the algorithm does not transwork from a stationario unit. It is showed to look at a relation assesses for the derivations at the above three from the shows the station of the its shows the static station of the static station is the showed terms from stationaries at the above.

<page-header><page-header><text><equation-block><equation-block><equation-block><equation-block><equation-block><text><equation-block><text><equation-block><text></text></equation-block></text></equation-block></text></equation-block></equation-block></equation-block></equation-block></equation-block></text></page-header></page-header>			
<text><equation-block><equation-block><equation-block><equation-block><text><equation-block><text><text><text></text></text></text></equation-block></text></equation-block></equation-block></equation-block></equation-block></text>			
<equation-block><equation-block><equation-block><equation-block><equation-block><text><equation-block><text><text></text></text></equation-block></text></equation-block></equation-block></equation-block></equation-block></equation-block>	52	CHAPTER 5. MAXIMUM LIKELIHOOD AND BAY	ESIAN ESTIMATION
<equation-block><equation-block><equation-block><text><equation-block><text><text><text><text></text></text></text></text></equation-block></text></equation-block></equation-block></equation-block>	f, which are given by		
<text><equation-block><text><text><text><text><text><text></text></text></text></text></text></text></equation-block></text>		$f(\theta) = \frac{1}{2} R(\theta) _2^2 = \frac{1}{2} \sum_{i=1}^{N} R_i(\theta)^2$	(5.36)
<text><equation-block><text><text><text><text><equation-block><text></text></equation-block></text></text></text></text></equation-block></text>		$\nabla f(\theta) = \sum_{i=1}^{N} \nabla R_i(\theta) R_i(\theta) = J(\theta)^{-} R(\theta)$	(5.37)
<equation-block><text><text><section-header><text><text></text></text></section-header></text></text></equation-block>		$\nabla^2 f(\theta) = \underbrace{J(\theta)^\top J(\theta)}_{-:B_{cin}(\theta)} + \sum_{i=1}^N \nabla^2 R_i(\theta) R_i(\theta)$	(5.38)
We note seen, as expected from a optimization algorithm, that the algorithm would not move away from a binary group with $V_1(h_0) = 0$. But the investor allow $L_1(h_0) = 0$, the two of $L_1(h_0) = 0$ is the two of $L_1(h_0) = 0$. The two of $L_1(h_0) = 0$ is the two of L_1	Using some of the above exp	ressions, the iterations of the Gauss-Newton algorithm can be	e written as
choose differently. If one would choose the inverse of the same providence of the solution of		$\boldsymbol{\theta}_{[k+1]} = \boldsymbol{\theta}_{[k]} - B_{\mathrm{GN}}(\boldsymbol{\theta}_{[k]})^{-1} \nabla f(\boldsymbol{\theta}_{[k]})$	
The noise ways variant a rough estimate of the conductors matrix $\Sigma_{i}^{(n)}$ of the matimate Γ rough be too match the the biodensities of the contrast σ for the discover in the term biodensities of the formation of the matimate Γ and σ	chosen differently. If one w the so-called Newton method class of so-called Newton repho- matrix $B_{\rm OW}(\theta)$ is called the but not necessarily positive Marquark algorithm, the G siep computation, for examp remain small enough for the Independent of which al a value θ^* that is an approx Le., set $\theta := \theta^*$. Interesting inverse Gauss-Newton Hessi	odd chose the inverse of the exact Hessian matrix, $V_{T}^{*}(\theta_{0}, W_{T}^{*}(\theta_{0}))$ descriptions of Hessian approximation gives note only of Gauss Gauss Network Hessian approximations. Note that it is a post- diary Network Hessian approximations for the compared by the new Network Hessian approximation is first compared by the the construct that the Hessian approximation for the optimization of the optimum Network and the second second second second second profit the study, at the end of the call of the optimization solution immede local minimizer of $f(\theta)$. We will use it as the maximiz- tion of the optimum Network and the optimization solution of the optimum Network and the second secon	() ¹¹ , one would obtain fifteent methods. The two semidefinite matrix, upple in the Levenberg- n modified in the actual definite or that the steps e actual function. er, the solver will return am-likelihood estimate, mpute afterwards – the
$\Sigma_{ij} = \frac{\ R(\theta^*)\ _{1}^{2}}{N-d} \left(P(\theta^*) \frac{1}{d}(\theta^*)\right)^{-1}$ If one wants to express the result of the parameter estimation, one often only uses the diagonal entries from this matrix, which contain, for $i = 1, \dots, d$, the variances σ_i^2 of the respective parameter components θ_i as seen in the following details derive sequences:	matrix Σ_c). Following the covariance matrix. Note th would be that the scale of th reason, the optimal squared have a <i>d</i> -dimensional vector discussed in Section 4.7. In noise covariance Σ_c , such th	linear least squares analysis, we could then directly use R_{i} at, due to the scaling in the expression $R(\theta) = \Sigma_{i}^{-1}(M(\theta)$ residual vector components is not only unitless, but also in th residual value is expected to be in the order of N . More precisi firing the data and minimizing the residual, we expect $ R(\theta)^{-1} _{C}$ and be different if practice, however, we might have made an error in estimating the size of the squared residual, $ R(\theta)^{-1} _{C}$ and be different if	$c_{DN}(\theta^*)^{-1}$ as parameter) = y), our assumption he order of one. For this ely, due the fact that we $ \frac{3}{2} \approx N - d$, as already the absolute size of the om $N - d$. Because this
If one wants to express the result of the parameter estimation, one often only uses the diagonal entries from this matrix, which contain, for $i = 1,, d$, the variances σ_i^2 of the respective parameter components $\hat{\theta}_{\mu}$ as seen in the following detailed matrix expression:			
	matrix, which contain, for i	result of the parameter estimation, one often only uses the dia = 1,, d_i the variances σ_i^2 of the respective parameter comp	agonal entries from this onents $\hat{\theta}_{is}$ as seen in the
[•••• a ²]	following detailed matrix ex		
		• • • σ_d^2	
	5.5. PRACTICAL SOLUT	ON OF THE NONLINEAR LEAST SQUARES PROBLEM	53
5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM 53	Taking the square root of t parameter estimation proced	he variances yields the standard deviations, such that the fi une could be presented by only 2d numbers in the form	nal result of the whole
5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM 53 Taking the square neet of the variances yields the standard deviations, such that the final result of the whole parameter estimation proceedure out by proceeding you by Januares in the form		$\theta_i = \theta_i^{\star} \pm \sqrt{\sigma_i^2}, \text{ for } i = 1, \dots, d$	
Taking the square root of the variances yields the standard deviations, such that the final result of the whole parameter estimation procedure could be presented by only 2d numbers in the form	5.5.3 Checking the C	ptimal Residual Vector	
Taking the square root of the variances yields the standard deviations, such that the final result of the whole parameter estimation procedure could be presented by only 2d numbers in the form $\theta_i = \theta_i^* \pm \sqrt{\sigma_i^2}, \text{for } i=1,\ldots,d$	Because the whole analysis will not be able to complete predictions for new situation not involved in the estimati	in this section is based on the measurement data that we use y answer the question of model validation, namely if our mode s. For model validation, we would need another set of meas on procedure, for example a new experiment that is perferent reviously conducted experiment that was kern secret during the	del is able to make valid surement data that were ned after the estimation

recorder and was just reserved for model validation. Bowvere, white we can do with the existing data of the single experiment that we use for parameter estimation, to look at the residual values $R_{i}(P)$ for i = 1, ..., N. If we plot them as a function of t, they should look singura of the residual values $R_{i}(P)$. The the binaryman look line the binaryman of the resonance of the singurant of the resonance of the resonance of the singurant of the sing

ML estimation

- For all possible parameter values:
 - Compute likelihood of the given observation(s)
- Pick the most likely!

Example problem



 Ferris wheel with (noisy) altimeters

Ferris Wheel



 $\sin(a) \approx -0.1$

 $\sin(a+30^\circ) \approx 0.6$

 $\sin(a + 60^\circ) \approx 0.9$

• Find the angle a !

Ferris Wheel



- $\sin(a) + \epsilon_1 = -0.1$ $\sin(a + 30^\circ) + \epsilon_2 = 0.6$ $\sin(a + 60^\circ) + \epsilon_3 = 0.9$
 - Assume that all ϵ are independent, and drawn from a normal distribution with zero mean and standard deviation 0.5
 - Find the most likely a !

```
% Defining a nonlinear model M
M = @(a) \sin(a + pi/6 * [0 1 2]')
% Our noise covariance matrix
C = diag(0.5^2 * [1 1 1])
% Square root of the covariance
S = sqrtm(C)
% Measurements
y = [-0.1, 0.6, 0.9]'
% The optimally weighted residual function
R = Q(a) inv(S) * (M(a) - y)
% Likelihood
% (During the lecure, I forgot the "0.5" below. Sorry about that! /Per)
p = @(a) exp(-0.5*sum(R(a).^2))
% A vector of values to try for the angle a
as = -pi:0.01:pi;
% Evaluate p for each element of the vector
% Note: Matlab automatically "broadcasts" opearations like exp, * and + across dimensions
ps = p(as);
% Alternatively, we could have written:
% ps = arrayfun(p, as);
% Another way would be:
% ps = zeros(size(as));
% for i=1:length(ps)
% ps(i) = p(as(i));
% end
% Plot p as funcition of a
plot(as, ps)
% Find the maximum and its location:
[pmax, ix] = max(ps)
a = as(ix)
```

5.5 Practical solution of the Nonlinear Least Squares Problem When we formulate and solve a nonlinear least squares problem, we need to use a numerical optimization routing to find the maximum likelihood estimate. In order to do this, we first scale the vector of model-measurement mismath residuals y = M(0) by using a - usually diagonal _geness Z₀ of the covariance marks of the noise, it 5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM order to obtain the scaled residual vector $R(\theta) := \Sigma_e^{-\frac{1}{2}} (M(\theta) - y)$ such that the maximum likelihood estimate $\hat{\theta} = \theta^*$ is obtained by the solution of the optimization problem $\theta^* = \arg \min_{\theta} \frac{\frac{1}{2} ||R(\theta)||_2^2}{-f(\theta)}$ (5.30) Note that the residual function R maps from \mathbb{R}^d to \mathbb{R}^N , and that most solution algorithms require the user to pass this function $R(\theta)$ – and not the objective function $f(\theta) = \frac{1}{2} |R(\theta)||_2^2$ – to the solver. We want to answer three constinuing this section: · How do we solve the nonlinear least squares optimization problem (5.30) in practice? · How can we obtain an estimate of the covariance matrix of the parameter estimate? · How can we assess if the modelling assumptions, in particular on the noise, were correct? We will answer the three questions in the following three subsections. 5.5.1 The Gauss-Newton Algorithm Solat the values version requiring the precisited solution of the presented of the precisited solution routines like MAT-LAB's legaron 11, which expect the user to provide an initial gases for the parameter $\theta =$ which we call θ_{ijk} . It is section – and a pointer to be functions $\mathbb{R}^2 \to \mathbb{R}^3$. Note validate lightlines are only guarantee to find local minima. Starting at the build gases θ_{ijk} , they, arrange a sequence of iterates that we call θ_{ijk} , θ_{ijk} , θ_{ijk} . Alter that each θ_{ijk} is a variet in the space \mathbb{R}^3 is \mathbb{R}^3 . Note that each θ_{ijk} is a variet in the space \mathbb{R}^3 is \mathbb{R}^3 . Note that θ_{ijk} is a variet in the space \mathbb{R}^3 is \mathbb{R}^3 . Note that θ_{ijk} is θ_{ijk} , θ_{ijk} . Alter index in the space requirement of all algorithms in that they head concepts to point θ_{ijk} that index is a start start of the so-called Gause-Newton method described next, though often there variants come during variations of the local distance \mathbb{R}^3 . Note that the order is the so-called Gause-Newton method described next, though often there variants come during variations of the local distance \mathbb{R}^3 is \mathbb{R}^3 . Note that \mathbb{R}^3 is \mathbb{R}^3 . Note \mathbb{R}^3 is \mathbb{R}^3 . The normalized \mathbb{R}^3 is \mathbb{R}^3 . Note \mathbb{R}^3 is \mathbb{R}^3 . Note \mathbb{R}^3 . asso we know very well how to solve linear losst squares problems and because all nonlinear functions to approximated by their first order Toylor series, a straightforward idea would be to solve a funct least form based on the functionariation at a sublucing most $k_{\pm 1}$ in ere to obtain a there to volve a function gases $k_{\pm 1}$ are solve to obtain $k_{\pm 1}$ within gases $k_{\pm 1}$ are due to the solve a function gase $k_{\pm 1}$ and $k_{\pm 1}$ are the solve of the first solve to find $k_{\pm 1}$ and $k_{\pm 1}$ and $k_{\pm 1}$ and $k_{\pm 1}$ are the first solve the first solve the first solve the solve first solve the solve of the solve the solve the first solve the $\theta_{[k+1]} = \arg\min_{\theta} \frac{1}{2} \left\| R(\theta_{[k]}) + \underbrace{\frac{\partial R}{\partial \theta}(\theta_{[k]})}_{\theta}(\theta - \theta_{[k]}) \right\|_{2}^{2}$ (5.31) $\begin{array}{l} = \arg\min_{\theta} \frac{1}{2} \left\| - J(\theta_{|k|}) \theta_{|k|} + R(\theta_{|k|}) + J(\theta_{|k|}) \theta \right\|_2^2 \\ = \left(J(\theta_{|k|})^\top J(\theta_{|k|})^{-1} J(\theta_{|k|})^{-1} J(\theta_{|k|}) \theta_{|k|} - R(\theta_{|k|}) \right) \\ = \theta_{|k|} - (J(\theta_{|k|})^\top J(\theta_{|k|}))^{-1} J(\theta_{|k|})^\top R(\theta_{|k|}) \\ = \theta_{|k|} - J(\theta_{|k|})^+ R(\theta_{|k|}) \end{array}$ (5.32) (5.33) (5.34) (5.35) ration above is only well defined if the Jacobian matrix $J(\theta_{|k|})$ is of full rank, but that in prac-tations, small algorithm modifications ensure that each iteration is well defined. With the above use already defined the basic Gauss-Neuton algorithm. One can show that -1 if converges – the algorithm converges linearly to a stationary point θ^* with $\nabla f(\theta^*) = 0$, but a proof of this result is understand the algorithm a bit better and to see at least why the algorithm does not move oint, it is useful to look at explicit expressions for the derivatives of the objective function CHAPTER 5. MAXIMUM LIKELIHOOD AND BAYESIAN ESTIMATION 52 f, which are given by $f(\theta) = \frac{1}{2} ||R(\theta)||_2^2 = \frac{1}{2} \sum_{i=1}^{N} R_i(\theta)^2$ (5.36) $\nabla f(\theta) = \sum_{i=1}^{N} \nabla R_i(\theta) R_i(\theta) = J(\theta)^\top R(\theta)$ (5.37) $\nabla^2 f(\theta) = \underbrace{J(\theta)^\top J(\theta)}_{i=1} + \sum_{i=1}^{N} \nabla^2 R_i(\theta) R_i(\theta)$ (5.38) isions, the iterations of the Gauss-Newton algorithm can be written as $\boldsymbol{\theta}_{[k+1]} = \boldsymbol{\theta}_{[k]} - B_{\mathrm{GN}}(\boldsymbol{\theta}_{[k]})^{-1} \nabla f(\boldsymbol{\theta}_{[k]})$ $B_{[n+1]} = B_{[n]} - B_{(n)}(B_{[n]}) \cdot \nabla T(B_{[n]})$ en, as expected from an orpinatization algorithm, but the algorithm would not move away from a first with $\nabla T(B_{[n]}) = 0$. But the inverted matrix $B_{(n)}(B_{[n]})^{-1}$ in from of the gradient could also be entropy. If one would choose the inverse of the exatt Hessin matrix $T(F_{(B_{[n]})})^{-1}$ and $T(B_{[n]})^{-1}$ model whether $T(B_{[n]})^{-1}$ models are also be associated with the exact the ex e., set $\hat{\theta} := \theta^*$. In 5.5.2 Estimating the Covariance Matrix and Extracting its Relevant Entries The easiest way to obtain a rough estimate of the covariance matrix Σ_{ϕ} of the parameter estimate θ^{*} would be a assume that the linearization of R at the solution is the correct model, and that all the statistical assumptions we made in the formation of the function R were covered; i.e., that we indeed that discussion noise with covariance matrix Σ_{c}). Following the linear least squares analysis, we could then directly use $R_{c}(\theta^{*})^{-1}$ as parameter c_{AB} recoming we mean new spaces margine we commute uncarge the $D_{CS}(V)^{-1}$ are produced and the mean matrix. Note that, due to the scaling in the expression $[0] = \sum_{i=1}^{N-1} (M_i O_i^{-1})_{i=i}$ for a simulation for that the optimal squared means that the scale of V_i means precisively, one for a simulation the optimal squared means V_i and V_i means V_i and V_i means V_i and V_i means V_i and V_i and V_i means V_i means V_i and V_i and V_i means V_i and V_i and V_i means V_i and V_i In protocol, the size of the squared residual $||R(\theta^-)||_2^2$ can be different from N-d. Beca llow the reasoning of Section 4.7, and in practice use the parameter covariance esti $\Sigma_{\delta} := \frac{\|R(\theta^*)\|_2^2}{N-d} (J(\theta^*)^\top J(\theta^*))^{-1}$ If one wants to express the result of the parameter estimation, one often only uses the diagonal entries from this matrix, which contain, for i = 1, ..., d, the variances σ_i^2 of the respective parameter components $\hat{\theta}_{ii}$ as seen in the following detailed matrix expression: $\Sigma_{\tilde{g}} = \begin{bmatrix} \sigma_1^2 & \bullet & \bullet \\ \bullet & \sigma_2^2 & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \sigma_d^2 \end{bmatrix}$

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM

Taking the square root of the variances yields the standard deviations, such that the final result of the whole parameter estimation procedure could be presented by only 2d numbers in the form $\theta_i = \theta_i^* \pm \sqrt{\sigma_i^2}$, for i = 1, ..., d

5.5.3 Checking the Optimal Residual Vector

5.5.3 Checking the Optimum retrievants re-con-Breams the whole analysis in this section is based on the measurement data that we use for the estimation, we will note be able to completely answer the question of model validation, namely if our model is able to make valid predictions for new simulators. For model validation, we would need another set of measurement data that were not involved in the estimation procedure, for example a new equipments that is preferred after the estimation procedure is finished, or a previously conducted experiment that a kept secret during the parameter estimation procedure is an way in tervers of or model validation.

procedure and was just reserved for model validation. Indexverx, what we can do with the oxising date of the single experiment that we use for parameter estimation, s to look at the residual values $R_i(P^*)$ for i = 1, ..., N. If we plot them as a function of *i*, they should look is the a sequence of nano numbers. In our other check this in more details, one pycholy creates and plots a situation of the residual values $R_i(P^*)$. If the biasogram looks like the biasogram of a zero mean Gaussian with intravirance, the model assumptions on the system and nuises are likely to be correct. It not, some part of the modelling assumptions was probably worng. One should then think hard and dange the system or esistem model on result to parameter containty proceedure, based on the same data, but on a different model.

- We cannot try all a.
- There are infinitely many.
- Let's solve a least squares problem!

When we formulate and solve a nonlinear least squares problem, we need to use a numerical optimization routin to find the maximum likelihood estimate. In order to do this, we first scale the vector of model-measurement mismath residuals $y = M(\theta)$ by using a - usually diagonal - guess Σ_0 of the covariance matrix of the neise, i

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM

order to obtain the scaled residual vector $R(\theta) := \Sigma_n^{-\frac{1}{2}} (M(\theta) - y)$ such that the maximum likelihood esti $\hat{\theta} = \theta^*$ is obtained by the solution of the optimization problem

 $\theta^* = \arg \min_{\theta} \frac{\frac{1}{2} \|R(\theta)\|_2^2}{-\theta(\theta)}$

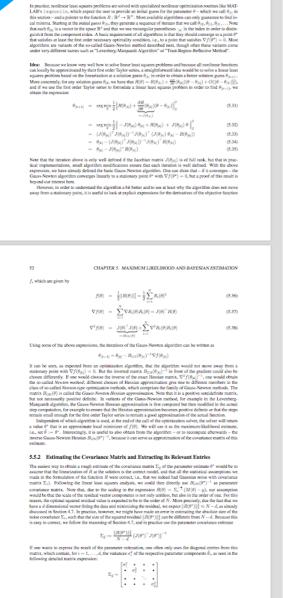
(5.30)

Note that the residual function R maps from \mathbb{R}^d to \mathbb{R}^N , and that most solution algorithms require the user to part this function $R(\theta)$ – and not the objective function $f(\theta) = \frac{1}{2} \|R(\theta)\|_2^2$ – to the solver. We want to answer thre questions in this section:

How do we solve the nonlinear least squares optimization problem (5.30) in practice?
 How can we obtain an estimate of the covariance matrix of the manmeter estimate?

How can we assess if the modelling assumptions, in particular on the noise, were correct?
 We will answer the three questions in the following three subsections,

5.5.1 The Gauss-Newton Algorithm



5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM

Taking the square root of the variances yields the standard deviations, such that the final result of the whole arameter estimation procedure could be presented by only 2d numbers in the form $\theta_i = \theta_i^* \pm \sqrt{\sigma_i^2}$, for i = 1, ..., d

5.5.3 Checking the Optimal Residual Vector

The concenting size spectra in the second second

as to look at the residual values $R(d^2)$ for i = 1, ..., N. If we plot them as a function of *i*, they should look due a sequence of moton numbers. In our bot to check this in mote edisis, one typically corresus adplots a situation of the residual values $R(d^2)$. If the biasegram looks like the biasegram of a zero mean Gaussian with intravirance, the mole assumptions on yeaves multiplicate the Ricky to be correct. How, some part of the modelling assumptions was probably swerge. One should then think hade and charge the system corises model on research models.

The Gauss-Newton method:

- Linearise the residual function at a point (your best guess)
- Solve a linear least squares problem (Find a better guess)
- Repeat!

 $\theta_{[k+1]} = \theta_{[k]} - J(\theta_{[k]})^{\dagger} R(\theta_{[k]})$

When we formulate and solve a nonlinear least squares problem, we need to use a numerical optimization routine to find the maximum likelihood estimate. In order to do this, we first scale the vector of model-measurement-mismatch residuals $y = \mathcal{M}(\theta)$ by using a – usually diagonal – guess Σ_v of the covariance matrix of the noise, in

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM

order to obtain the scaled residual vector $R(\theta) := \Sigma_c^{-\frac{1}{2}}(M(\theta) - y)$ such that the maximum likelihood estimate $\hat{\theta} = \theta^*$ is obtained by the solution of the optimization problem

 $\theta^* = \arg\min_{\boldsymbol{\theta}} \underbrace{\frac{1}{2} \|\boldsymbol{R}(\boldsymbol{\theta})\|_2^2}_{-f(\boldsymbol{\theta})}$

\$1

(5.30)

Note that the residual function R maps from \mathbb{R}^d to \mathbb{R}^N , and that most solution algorithms require the user to pass this function $R(\theta) - and not the objective function <math display="inline">f(\theta) = \frac{1}{2} \|R(\theta)\|_{1}^{2}$ – to the solver. We want to answer three questions in this section:

How do we solve the nonlinear least squares optimization problem (5.30) in practice?

· How can we obtain an estimate of the covariance matrix of the parameter estimate? · How can we assess if the modelling assumptions, in particular on the noise, were correct?

We will answer the three questions in the following three subsections.

5.5.1 The Gauss-Newton Algorithm

2.5.1 LTPC VBBBS-VEWON Approximate In practice, nonlinear loss sparses problems are volved with specialized notificater optimization nontines like MAT-LAPS 1 appon11m, which report the user to provide an initial guess for the parameter $\theta =$ which we call θ_{gin} that we like a specificate on the loss of the maximum start of the laplacet of laplacet of the laplacet of the laplacet of the laplacet of laplacet of the laplacet of laplacet of laplacet of the laplacet of laplacet

Idea: Because we know very well how to solve linear least squares problems and because all nonlinear functions can locally be approximated by their first order Taylor entres, a straightforward idea would be to solve a linear least squares problem based on the linearization at solution gases δ_{+1} , in order to obtain a their solution gases δ_{+1} . More concernedy, for any solution gases δ_{+1} , we have that $\mathcal{M}(\theta) = \mathcal{M}(\theta_{+}) = \mathcal{M}(\theta_{-}) = \delta_{+1}(\theta_{-}) = \delta_{+1}(\theta_{-})$.

 $\theta_{[k+1]} \hspace{.1 in} = \hspace{.1 in} \arg\min_{\boldsymbol{\vartheta}} \frac{1}{2} \left\| R(\boldsymbol{\theta}_{[k]}) + \underbrace{\frac{\partial R}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_{[k]})}_{\boldsymbol{\vartheta}} (\boldsymbol{\theta} - \boldsymbol{\theta}_{[k]}) \right\|_2^2$ (5.31) $\begin{array}{rl} & -J(n_{i}) \\ = & \arg\min_{k} \frac{1}{2} \Big\| - J(\theta_{[k]}) \, \theta_{[k]} + R(\theta_{[k]}) + J(\theta_{[k]}) \, \theta \Big\|_{2}^{2} \\ = & (J(\theta_{[k]})^{T} J(\theta_{[k]}))^{-1} J(\theta_{[k]})^{-1} J(\theta_{[k]}) - R(\theta_{[k]}) \\ = & \theta_{[k]} - (J(\theta_{[k]})^{T} J(\theta_{[k]}))^{-1} J(\theta_{[k]})^{-1} R(\theta_{[k]}) \\ = & \theta_{[k]} - J(\theta_{[k]})^{-R} R(\theta_{[k]}) \end{array}$ (5.32)

(5.33) (5.34) (5.35)

Note that the iteration above is only well defined if the absorban matrix $J(\theta_{(0)})$ is of full rank, but that in practical implementations, small algorithm modifications ensure that each bratica is swell defined. With the above expression, we have easily defined the balac Gauss-Newton algorithm coverges - the Gauss-Newton algorithm coverges interval to a stationary point θ^{*} with $\nabla(\theta^{*}) \rightarrow 0$, but a proof of this result is byound or interval better. However, is order to understand the algorithm a bit better and to see all as why the algorithm does not move say from a stationary point. It is well to be size a regular expressions for the derivations of the object function and the point of the state of t

$ \begin{array}{c} \sum_{i=1}^{2} \nabla R_i(\theta) = \sum_{i=1}^{2} \nabla R_i(\theta) R_i(\theta) = J(\theta)^{-1} R(\theta) \qquad (5) \\ \nabla^2 f(\theta) = \int_{0}^{1} \int_{0}^{1} J(\theta) + \sum_{i=1}^{N} \nabla^2 R_i(\theta) R_i(\theta) \qquad (5) \\ \nabla^2 f(\theta) = \int_{0}^{1} \int_{0}^{1} J(\theta) + \sum_{i=1}^{N} \nabla^2 R_i(\theta) R_i(\theta) \qquad (5) \\ \sum_{i=1}^{N} (\theta) R_i(\theta) + \sum_{i=1}^{N} \nabla^2 R_i(\theta) R_i(\theta) \qquad (5) \\ \sum_{i=1}^{N} (\theta) R_i(\theta) + \sum_{i=1}^{N} \nabla^2 R_i(\theta) R_i(\theta) \qquad (5) \\ \sum_{i=1}^{N} (\theta) R_i(\theta) + \sum_{i=1}^{$	52	CHAPTER 5. MAXIMUM LIKELIHOOD AND BAYE	SIAN ESTIMATI
$\nabla f(\theta) = \sum_{i=1}^{N} \nabla R_i(\theta) R_i(\theta) = 0/ \theta ^2 R_i(\theta) = 0$ $(\zeta - Q)^2 + (\zeta - Q) = \sum_{i=1}^{N} \nabla R_i(\theta) R_i(\theta) = 0/ \theta ^2 - R_i(\theta) = 0$ (3) $\nabla^2 f(\theta) = \frac{1}{2-\alpha_{i+1}(\theta)} + \sum_{i=1}^{N} \nabla^2 R_i(\theta) R_i(\theta) = 0$ Using some of the above expression, the iterations of the Gauss-Newton algorithm can be written as $\theta_{N+1} = \theta_{N-1} - \theta_{N-1} - \theta_{N-1} - \theta_{N-1} = 0$ (4) Is one bases, as expected from an equivalent algorithm. Such as <i>dga</i> -thm would not move away for stationary point with $\nabla_i^2(\theta_N) = 0$. But the inversed matrix $R_{i+1}^2(\theta_N) = 1$ is front of the gaussian could also the so-called Newton-type equivalent methods, which comprises the flamity of Gauss-Newton algorithm, $R_{i+1}^2(\theta_N) = 0$. But the inversed matrix $R_{i+1}^2(\theta_N) = 0$ must be coulded Newton endowed, after methods, matrix $R_{i+1}^2(\theta_N) = 0$. The the inverse of matrix $R_{i+1}^2(\theta_N) = 0$ must be coulded Newton endowed frame combestions of the state and such that the Gauss-Newton algorithm is denoted in the two scale of Newton endowed Neural Hermiterian denotes in the transmitteriant $R_{i+1}^2(\theta_N) = 0$ and the Hermiterian equivalent the optimum of the data interval the state of the state of the data interval the state of the st	f, which are given by		
$\begin{split} & \sum_{k=0}^{n-1} (\theta_k) = d(1)^{-1} f(\theta_k) + \sum_{k=1}^{N} \nabla^2 R_k(\theta) R_k(\theta) (6) \\ & (5) \\ &$		$f(\theta) = \frac{1}{2} R(\theta) _2^2 = \frac{1}{2} \sum_{i=1}^{N} R_i(\theta)^2$	(5.
Using some of the above expressions, the iteration of the Gauss-Newton algorithm can be written as $\theta_{\beta+1 } = \theta_{\beta } - B_{CN}(\theta_{\beta })^{-1} \nabla f(\theta_{\beta })$. It can be seen, as expected from an optimization algorithm, that the algorithm would not more away for statismary point with $V(\theta_{\beta }) = 0$. But the inverte of the scale $S(\theta_{\beta })^{-1}$ is find of the gradient could above all directly. If one would choose the inverte of the scale $S(\theta_{\beta })^{-1}$ is not of the gradient could be scaling blockous models, different choices of Hussian approximation joy enter set of them rambers in matrix $B_{2N}(p)$ is called the Gauss-Nerone Hessian approximation. Note that it is a positive semidations the number of the scale of the scale approximation is first compared but then modified in the ac- size comparison (or incomplete densities). The Hessian approximation becomes positive definite. In some first and mough for the first order Taylor set is tremain a good approximation is the actual formation. Strengther with the scale of the scale approximation is the cample in the Level scale of "the its are approximated to the lines in approximation is the cample in the Level matrix $B_{2N}(p)$ is called the Gauss-Nerone Hessian approximation is first compared but then modified in the ac- size comparison (or in the lines in approximation is first compared but them modified in the ac- size comparison (or in the Hessian approximation is first compared but them modified in the ac- size of θ_{2N} is a scale is to also obtain from the algorithm – or to recompare direvanch- interes Gauss-Nerone Hessian $P_{2N}(p)$. The scale is a stress the accurate matrix to the matrix $B_{2N}(p)$ is a scale in the above the scale approximation of the actual direvanch- interest Gauss-Nerone Hessian $B_{2N}(p)$. The scale is a stress the approximation of the actual direvanch- interest Gauss-Nerone Hessian $B_{2N}(p)$. The scale is a stress the approximation of the actual direvanchesis and the direct is the scale approximation in the		$\nabla f(\theta) = \sum_{i=1}^{N} \nabla R_i(\theta) R_i(\theta) = J(\theta)^{-} R(\theta)$	(5.
$\theta_{\beta-1} = \theta_{\beta_0} - B_{(2)}(\theta_{\beta_0})^{-1} \nabla f(\theta_{\beta_0})$ Is can be seen, as expected from an optimization algorithm, that the digerithm would not more away for substraney point with $Q(\theta_{\beta_0}) = 0$. But the inverse of the scale (Bessian matrix, $\nabla^2 f(\theta_{\beta_0})^{-1}$, one would choose the inverse of the scale (Bessian matrix, $\nabla^2 f(\theta_{\beta_0})^{-1}$, one would choose the inverse of the scale (Bessian matrix, $\nabla^2 f(\theta_{\beta_0})^{-1}$, one would choose the inverse of the scale (Bessian matrix, $\nabla^2 f(\theta_{\beta_0})^{-1}$, one would choose the inverse of the scale (Bessian matrix). The scale of the scale (Bessian matrix) and (Bessian matrix), the scale (Bessian matrix), the scale (Bessian matrix), the scale (Bessian matrix), the scale (Bessian Messian (Bessian Bessian		$\nabla^2 f(\theta) = \underbrace{J(\theta)^\top J(\theta)}_{-:\mathcal{B}_{CN}(\theta)} + \sum_{i=1}^N \nabla^2 R_i(\theta) R_i(\theta)$	(5.
It can be seen, as expected from a solution all approximations and the disputches would use more using for assuming point with $V(f_0)_{(1)} = 0$ for the invested miniter, $V(f_0)_{(1)} = 0$ minites of the predict evolution chosen differently. It one would choose the investe of the start Beasian matrix, $V(f_0)_{(1)} = 1$, now would of the so-called Neuron works, different choices of Hessian approximation prior network works would be the so-called Neuron system springiation methods, which comprises the future of the start Beasian methods, ministra $R_{(0,1)}(0)$ and the Gauss Neuron Hessian approximations. Nest that it is a positor semidiative mo- thon starts $R_{(0,1)}(0)$ accounds the Gauss Approximation is first compared by them motion of matrix $R_{(0,1)}(0)$ accounds to some the Hessian approximation in first compared by them motion of the start approximation of the instart of the start of the start of the start of the start first of the start theory is a start of the start first of the start is used to be approximation in the start of the start of the start of the start theorem in the start of the start is the start of the start is the start of the start is the start of the start is the start of the start is the start of the start is the start of the start is the start of the start is the start of the start of the start of the start of the s	Using some of the abo	ove expressions, the iterations of the Gauss-Newton algorithm can be w	written as
stationary point with $\nabla_{1}^{1}(\theta_{01}) \rightarrow 0$. But the inverted matrix $B_{col}(\theta_{01})^{-1}$ in front of the gradient could also chosen differently. How would choose the inverte of the acut behavior matrix $\nabla_{1}^{1}(\theta_{01})^{-1}$, one would the facts of the would behavior the inverted of the acut behavior matching $\nabla_{1}^{1}(\theta_{01})^{-1}$, one would choose the matrix $B_{col}(\theta_{01})$ statistical due factors of Hessian approximation is prior into the different members in class of so-called wetters reproduced by the inverted of the source o		$\theta_{[k+1]} = \theta_{[k]} - B_{GN}(\theta_{[k]})^{-1} \nabla f(\theta_{[k]})$	
	stationary point with chosen differently. If the so-called Newton class of so-called New matrix $B_{GN}(\theta)$ is call but not necessarily p Manquarkt algorithm, step computation, for remain small enough Independent of w a value θ^* that is an a i.e., set $\theta := \theta^*$. Inte	$\nabla_{i}^{j}(\theta_{ij}\rangle) = 0$. But the inverse of matrix $R_{ij}(\theta_{ij}\rangle)^{-1}$ in front of the moves of the exact Hessian matrix, $\nabla_{i}^{j}(\theta_{ij}\rangle)$ method, different choices of Hessian approximation give rules to different choices of Hessian approximation, Nort fund it is a positive of the Gauss-Neutron Hessian approximation is fract compated but then example to ensure that the Hessian approximation is fract compated but then example to ensure that the Hessian approximation is fract compated but then example to ensure that the Hessian approximation is fract compated but then example to ensure that the Hessian approximation is fract compated but then example to ensure that the Hessian approximation is fract on the start approximate local minimizer of $j(\theta)$. We will use it us the maximum ensuingly, at is welft to also obtain from the algorithm — or to recom-	pradient could also pradient could also ierent members in Newton methods. 1 e semidefinite mat ple in the Levenbe modified in the act finite or that the st actual function. the solver will ret n-likelihood estim pute afterwards –
	5.5.2 Estimating	g the Covariance Matrix and Extracting its Relevant Er	itries

The using using to obtain a rough estimate of the containen matrix Σ_{ij} of the parameter estimate T would be to summa that the household of the distribution of the control and shared that the matriceal assumptions we made in the formulation of the function R were correct, i.e., that we indeed that Giussian noise with covariance matrix Σ_{ij} . Following the limits characterized seques studys, we could then discretize $R(R)=N_{ij}^{-1}(R(R)-R_{ij})$, one assumption with Σ_{ij} . Following the limits characterized constraints $R(R)=N_{ij}^{-1}(R(R)-R_{ij})$, one assumption with the original sequestion $R(R)=N_{ij}^{-1}(R(R)-R_{ij})$, one assumption were a dominant order for the study and matrix initial performance in sections R(R). The sum of the fact that we are dominant order for the study of the summation R(R) is a similarly discovered in Section 4.7. In practice, however, we might have made as zero in estimating the should us that R_{ij} sections R(R) is a source of the summation of the summation R(R) is a source of the summation of the summation R(R) is a source of the summation of the summation R(R) and R(R) is a source of the summation of the summation R(R) is a source of the summation of the summation R(R) and R(R) is a source of the summation of the summation R(R) and R(R) is a source of the summation of the summation R(R) and R(R) is a source of the summation of the summation R(R) and R(R) is a source of the summation of the summation R(R) and R(R) is a source of the summation of the summation R(R) and R(R) is a source of the summation of the summation R(R) and R(R) is a source of the summation of the summation R(R) is a source of the summation R(R) and matrix R(R) and R(R) is a source of the summation R(R) and R(R) are defined as the source of the summation R(R) and R(R

 $\Sigma_{\delta} := \frac{\|R(\theta^*)\|_2^2}{N-d} \left(J(\theta^*)^\top J(\theta^*)\right)^{-1}$ If one wants to express the result of the parameter estimation, one often only uses the diagonal entries from this matrix, which contain, for $i = 1, ..., d_i$ the variances σ_i^2 of the respective parameter components θ_i , as seen in the following detailed matrix expression:



5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM 53

Taking the square root of the variances yields the standard deviations, such that the final result of the whole parameter estimation procedure could be presented by only 2d numbers in the form $\theta_i = \theta_i^* \pm \sqrt{\sigma_i^2}$, for i = 1, ..., d

5.5.3 Checking the Optimal Residual Vector

5.5.3 Checking the Optimal Residual Vector
Because the whole analysis in this sector is hard on the measurement data that we use for the estimation, we will not be able to completely answer the agestion of model validation, analysis of this sector that we not involved in the estimation procedure, for example, a new caperiment that is performed after the estimation procedure in finishes (are a provide) validation, we waited the sector that we not involved in the estimation procedure in finishes (are provide) validation, we waited the sector that we not involved in the simulation procedure in finishes (are provide) validation, we waited the sector and was a performed after the estimation procedure and was just reserved for model validation.
However, what we can be with the estimation galaxies that have keep seter disting the parameter estimation procedure and was just reserved for model validation.
However, what we can be with the estimation galaxies that have have the estimation of the set of the sector of the set of the sector disting data of the site is the sector of the sector and processing of the sector o

$$egin{aligned} heta_{[k+1]} &= & rg\min_{ heta} rac{1}{2} \Big\| R(heta_{[k]}) + \underbrace{rac{\partial R}{\partial heta}(heta_{[k]})}_{=:J(heta_{[k]})} (heta - heta_{[k]}) \Big\|_2^2 \end{aligned}$$

$$= \arg\min_{\theta} \frac{1}{2} \left\| -J(\theta_{[k]}) \ \theta_{[k]} + R(\theta_{[k]}) \ + \ J(\theta_{[k]}) \ \theta \right\|_2^2$$

$$= (J(\theta_{[k]})^{\top}J(\theta_{[k]}))^{-1}J(\theta_{[k]})^{\top} (J(\theta_{[k]}) \ \theta_{[k]} - R(\theta_{[k]}))$$

$$= \theta_{[k]} - (J(\theta_{[k]})^{\top} J(\theta_{[k]}))^{-1} J(\theta_{[k]})^{\top} R(\theta_{[k]})$$

Note: We do not use the Hessian of R

When we formulate and solve a nonlinear least squares problem, we need to use a numerical optimization routine to find the maximum likelihood estimate. Is order to do this, we first scale the vector of model-neasurement-mismath residuals $y = M(\theta)$ by using a - sousily diagonal - goess x_i of the covariance matrix of the solve, in

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM

order to obtain the scaled residual vector $R(\theta) := \Sigma_c^{-\frac{1}{2}} (M(\theta) - y)$ such that the maximum likelihood estimate $\hat{\theta} = \theta^*$ is obtained by the solution of the optimization problem

 $\theta^* = \arg \min_{\theta} \frac{\frac{1}{2} \|R(\theta)\|_2^2}{-f(\theta)}$

51

(5.30)

Note that the residual function R maps from \mathbb{R}^d to \mathbb{R}^N , and that most solution algorithms require the user to pass this function $R(\theta) = ad$ not the objective function $f(\theta) = \frac{1}{2} \|R(\theta)\|_{1}^{2}$ —to the solver. We want to answer three questions in this section:

· How do we solve the nonlinear least squares optimization problem (5.30) in practice? · How can we obtain an estimate of the covariance matrix of the parameter estimate?

· How can we assess if the modelling assumptions, in particular on the noise, were correct?

We will answer the three questions in the following three subsections.

5.5.1 The Gauss-Newton Algorithm

The consists external region and provide the set of th

Idea: Because we know very well how to solve linear least squares problems and because all nonlinear functions can locally be approximated by their first order. Taylor series, a straightforward dise would be to solve a linear least the other concernels, for any solutions games by use how that $R(0) = R(\theta_0) = \frac{100}{2} (\theta_0 - \theta_0) = \frac{10$

 $\theta_{[k+1]} = \arg\min_{\theta} \frac{1}{2} \left\| R(\theta_{[k]}) + \underbrace{\frac{\partial R}{\partial \theta}(\theta_{[k]})}_{2}(\theta - \theta_{[k]}) \right\|_{2}^{2}$ (5.31) (5.32)

$$\begin{split} & - \langle \theta_{(k)} \rangle \\ = & \arg \min_{k} \frac{1}{2} \left\| - J(\theta_{(k)}) \theta_{(k)} + R(\theta_{(k)}) + J(\theta_{(k)}) \theta \right\|_{2}^{2} \\ & - (J(\theta_{(k)})^{-1}J(\theta_{(k)})^{-1}J(\theta_{(k)})^{-1}J(\theta_{(k)}) \theta_{(k)} - R(\theta_{(k)})) \\ & - \theta_{(k)} - (J(\theta_{(k)})^{-1}J(\theta_{(k)})^{-1}J(\theta_{(k)})^{-1}R(\theta_{(k)}) \\ & - \theta_{(k)} - J(\theta_{(k)})^{-1}R(\theta_{(k)}) \end{split}$$
(5.33) (5.34) (5.35)

Note that the iteration above is only well defined if the laceobian matrix $J(\theta_{[N]})$ is of full rank, but that in practical implementations, small algorithm modificatione susue that each iteration is well defined. With the above previous, we have alimptic dimensional structures and the dimensional structure of the structure

CHAPTER 5. MAXIMUM LIKELIHOOD AND BAYESIAN ESTIMATION 52 f, which are given by

> $f(\theta) = \frac{1}{2} ||R(\theta)||_2^2 = \frac{1}{2} \sum_{i=1}^{N} R_i(\theta)^2$ (5.36) $\nabla f(\theta) = \sum_{i=1}^{N} \nabla R_i(\theta) R_i(\theta) = J(\theta)^\top R(\theta)$ (5.37) $\nabla^2 f(\theta) = \underbrace{J(\theta)^{\top}J(\theta)}_{=:\mathcal{B}_{cN}(\theta)} + \sum_{i=1}^N \nabla^2 R_i(\theta) R_i(\theta)$ (5.38)

Using some of the above expressions, the iterations of the Gauss-Newton algorithm can be written as

 $\theta_{[k+1]} = \theta_{[k]} - B_{GN}(\theta_{[k]})^{-1}\nabla f(\theta_{[k]})$

 $\begin{array}{l} \theta_{g+1} = \theta_g - \Omega_{g-1}(\theta_g) \cdot \nabla_{g-1}(\theta_g) (\theta_g) \cdot \nabla_{g-1}(\theta_g) \\ \text{is can be seen, so repected from an optimization algorithm, but the algorithm would not move away from a stationary point with <math display="inline">\nabla(\theta_{g-1}) = 0.$ But the inverse of matrix $B_{g-1}(\theta_g) \cdot 1$ in front of the gradient could also be chosen differently. To see would chosen the inverse of the actual Heasian matrix, $\nabla(\theta_{g-1}) -$ one would obtain the inverse of the actual Heasian matrix, $\nabla(\theta_{g-1}) -$ one would obtain the inverse of the actual Heasian matrix, $\nabla(\theta_{g-1}) -$ one would obtain the inverse of the actual Heasian matrix, $\nabla(\theta_{g-1}) -$ one would obtain the inverse of the actual state of the entry of the actual state B_{g-1} and the second Heasian Heasian

5.5.2 Estimating the Covariance Matrix and Extracting its Relevant Entries

5.5.2 Holimating the Covariance Matrix and Extracting its Kelevant Entries The existent syst oxinia norugh entimise of the covariance mutitix Σ_0 of the parameter estimate θ^* would be to assume that the linearization of R at the solution is the correct model, and that all the statistical assumptions we made in the formation of the function N preveres exercit, e. that we indeed had Gaussian noise with covariance matrix Σ_0 . Following the linear least squares analysis, we could then directly use $B_{CN}(\theta^*)^{-1}$ is parameter to assume that the scale of the residual vector components is not only unlikes, but also in the outer of one. For this means, the optimal squared residual values is expected to be in the outer of N. More precisely, during that that and have a dimensional vector fitting the data and minimizing the residual, we repect $[R(\theta^*)]_{11}^{-1} = N - L_{-n}^{-1}$ is intrapied weak constructions. Work the the vector is the squared residual ($R(\theta^*)$) is constructions for $N - L_{-n}$ is intrapied basic covariance. This, which the viscine is the squared residual ($R(\theta^*)$) is constructions from N - d. Bosome this is easy to correct, we follow the reasoning of Section 4.7, and in practice use the parameter covariance estimate

 $\Sigma_{\hat{\theta}} := \frac{\|R(\theta^*)\|_2^2}{N-d} \left(J(\theta^*)^\top J(\theta^*)\right)^{-1}$

If one wants to express the result of the parameter estimation, one often only uses the diagonal entries from this matrix, which contain, for i = 1, ..., d, the variances σ_i^2 of the respective parameter components $\hat{\theta}_{ii}$ as seen in the following detailed matrix expression

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM 53

Taking the square root of the variances yields the standard deviations, such that the final result of the whole parameter estimation procedure could be presented by only 2*d* numbers in the form $\theta_i = \theta_i^* \pm \sqrt{\sigma_i^2}$, for i = 1, ..., d

5.5.3 Checking the Optimal Residual Vector

5.5.3 Checking the Optimal Residual Vector
Because the whole analysis in this sector is hard on the measurement data that we use for the estimation, we will not be able to completely answer the agestion of model validation, analysis of this sector that we not involved in the estimation procedure, for example, a new caperiment that is performed after the estimation procedure in finishes (are a provide) validation, we waited the sector that we not involved in the estimation procedure in finishes (are provide) validation, we waited the sector that we not involved in the simulation procedure in finishes (are provide) validation, we waited the sector and was a performed after the estimation procedure and was just reserved for model validation.
However, what we can be with the estimation galaxies that have keep seter disting the parameter estimation procedure and was just reserved for model validation.
However, what we can be with the estimation galaxies that have have the estimation of the set of the sector of the set of the sector disting data of the site is the sector of the sector and processing of the sector o

$$\begin{split} f(\theta) &= \frac{1}{2} \|R(\theta)\|_2^2 = \frac{1}{2} \sum_{i=1}^N R_i(\theta)^2 \\ \nabla f(\theta) &= \sum_{i=1}^N \nabla R_i(\theta) R_i(\theta) = J(\theta)^\top R(\theta) \\ \nabla^2 f(\theta) &= \underbrace{J(\theta)^\top J(\theta)}_{=:B_{\rm GN}(\theta)} + \sum_{i=1}^N \nabla^2 R_i(\theta) R_i(\theta) \end{split}$$

$$\theta_{[k+1]} = \theta_{[k]} - B_{\text{GN}}(\theta_{[k]})^{-1} \nabla f(\theta_{[k]})$$

$\theta = [a]$

```
% The Jacobian of the R function:
J = @(a) inv(S) * cos(a + pi/6 .* [0 1 2]')
% Iterations of the Gauss-Newton algoritm:
a = a - J(a) \ R(a)
% ...after a few iterations we should have a good estimate of the
% argmin of norm(R(a))^2
```

When we formulate and solve a nonlinear least squares problem, we need to use a numerical optimization routine to find the maximum likelihood estimate. Its order to do fits, we first scale the vector of model-measurement-mismath residuals $y = M(\theta)$ by using a - sound y diagonal - goess y, of the covariance matrix of the noise, in

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM

order to obtain the scaled residual vector $R(\theta) := \Sigma_c^{-\frac{1}{2}}(M(\theta) - y)$ such that the maximum likelihood estimate $\hat{\theta} = \theta^*$ is obtained by the solution of the optimization problem

 $\theta^* = \arg \min_{\theta} \frac{\frac{1}{2} \|R(\theta)\|_2^2}{-f(\theta)}$

\$1

(5.30)

Note that the residual function R maps from \mathbb{R}^d to \mathbb{R}^N , and that most solution algorithms require the user to pass this function $R(\theta) = and$ not the objective function $f(\theta) = \frac{1}{2} \|R(\theta)\|_{1}^2$ – to the solver. We want to answer three questions in this section:

· How do we solve the nonlinear least squares optimization problem (5.30) in practice? · How can we obtain an estimate of the covariance matrix of the parameter estimate?

· How can we assess if the modelling assumptions, in particular on the noise, were correct?

We will answer the three questions in the following three subsections.

5.5.1 The Gauss-Newton Algorithm

2.3.1 The values' version rapping in the present set work with specialized perimeter optimization routines like MAT-LAB's lapponlin, which expect the user to provide an initial gases for the parameter $\theta =$ which we call θ_{gin} that its section - and a point no test factors for $\mathbb{R}^{d} \to \mathbb{R}^{d}$. Note valid that $\mathbb{R}^{d} \to \mathbb{R}^{d}$. Note that $\mathbb{R}^{d} \to \mathbb{R}^{d}$. The factor is the specific different is a specific different in the specific different is a specific different in the specific different is different in the specific different in the specific different is different in the specific different is different in the specific different is different in the specific different in the specific different is different in the specific different in the specif

Idea: Because we know very well how to solve linear least squares problems and because all nonlinear functions can locally be approximated by their first order Taylor series, a straightforward data would be to solve a linear least square problem saved on the linearization at solution graves δ_{ij} , in order to data a lenstre solution graves δ_{ij} , More concrutely, for any solution graves δ_{ijk} , we have that $R(0) - R(\theta_k) + \frac{R(\theta_k) - R(\theta_k) - R(\theta_k)}{R(\theta_k) - R(\theta_k) - R(\theta_k)}$.

 $\theta_{[k+1]} = \arg\min_{\theta} \frac{1}{2} \left\| R(\theta_{[k]}) + \underbrace{\frac{\partial R}{\partial \theta}(\theta_{[k]})}_{0} (\theta - \theta_{[k]}) \right\|_{2}^{2}$ (5.31) (5.32)

 $\begin{array}{rl} & -J(\theta_{kl}) \\ = & \arg\min_{k} \frac{1}{2} \left\| -J(\theta_{kl}) \, \theta_{[k]} + R(\theta_{kl}) + J(\theta_{[k]}) \, \theta \right\|_{2}^{2} \\ = & (J(\theta_{[k]})^{T} J(\theta_{[k]}))^{-1} J(\theta_{[k]})^{T} (J(\theta_{[k]}) \, \theta_{[k]} - R(\theta_{[k]})) \\ = & \theta_{[k]} - (J(\theta_{[k]})^{T} J(\theta_{[k]}))^{-1} J(\theta_{[k]})^{T} R(\theta_{[k]}) \\ = & \theta_{[k]} - J(\theta_{[k]})^{T} R(\theta_{[k]}) \end{array}$ (5.33) (5.34) (5.35)

Note that the iteration above is only well defined if the also backing matrix $J(\theta_{[0]})$ is of full rank, but that in practical implementations, small algorithm modifications ensure that each lettacion is well defined. With the above expression, we have easily defined the back Gauss-Newton algorithm. Once on how that -1 if it converges - the Gauss-Newton algorithm coverges) interfy to a stationary point θ^* with $\nabla (|\theta^*) = -0$, but a proof of this result is byound our interest. Here, it is obtained the algorithm of the state of the interest is a start why the algorithm does not move say from a stationary point, it is useful to look at explicit carpensions for the deviatives of the elected function.

CHAPTER 5. MAXIMUM LIKELIHOOD AND BAYESIAN ESTIMATION 52 f, which are given by

 $f(\theta) = \frac{1}{2} ||R(\theta)||_2^2 = \frac{1}{2} \sum_{i=1}^{N} R_i(\theta)^2$ (5.36) (5.37) $\nabla f(\theta) = \sum_{i=1}^{N} \nabla R_i(\theta) R_i(\theta) = J(\theta)^\top R(\theta)$ $\nabla^2 f(\theta) = \underbrace{J(\theta)^\top J(\theta)}_{==0} + \sum_{i=1}^N \nabla^2 R_i(\theta) R_i(\theta)$ (5.38)

Using some of the above expressions, the iterations of the Gauss-Newton algorithm can be written as

 $\theta_{[k+1]} = \theta_{[k]} - B_{GN}(\theta_{[k]})^{-1} \nabla f(\theta_{[k]})$

 $\begin{array}{l} \theta_{p,11} = \theta_{p,-1} - \theta_{p,-2} - \theta_{p,-2} - \theta_{p,-1} - \theta_{p,-1} + \theta_{p,-1} - \theta_{p$

5.5.2 Estimating the Covariance Matrix and Extracting its Relevant Entries

5.2. E hstimating the Covariance Matrix and Extracting in Keievant Entires The existent way to obtain a rough estimate of the covariance matrix Σ_0 of the parameter estimate θ^* would be to assume that the linearization of R at the vehation is the correct model, and that all the statistical assumptions we and in the formation of the function W were correct, i.e., that we indeed that Gaussian noise with covariance matrix Σ_0 . Following the linear least squares analysis, we could then directly use $B_{CS}(\theta^*)^{-1}$ is parameter would be that the scale of the residual vector component is not only unlikes, but also in the outer of one. For this masso, the optimal square residual value is expected to be in the outer of N. More precisely, due the fact that are here a d-dimensional vector fitting the date and maintaining the residual, we report $[R(\theta^*)]_{0}^{-1} = N - L_{-n}^*$ as larged particle covariance. Such that the vise of the squared residual ($[R(\theta^*)]_{0}^{-1}$ can be covered by the scale of the residual vector fitting the date and maintaining the residual ($[R(\theta^*)]_{0}^{-1}$ can be covered by the scale of the residual vector fitting the date and maintaining the residual vector fitting the date and maintain the residual vector fitting the date and maintain the term of the date and maintain the residual vector fitting the

 $\Sigma_{\hat{\theta}} := \frac{\|R(\theta^*)\|_2^2}{N-d} \left(J(\theta^*)^\top J(\theta^*)\right)^{-1}$ If one wants to express the result of the parameter estimation, one often only uses the diagonal entries from this matrix, which contain, for i = 1, ..., d, the variances σ_i^2 of the respective parameter components θ_{ii} as seen in the following detailed matrix expression.



5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM 53

Taking the square root of the variances yields the standard deviations, such that the final result of the whole parameter estimation procedure could be presented by only 2d numbers in the form $\theta_i = \theta_i^* \pm \sqrt{\sigma_i^2}$, for i = 1, ..., d

5.5.3 Checking the Optimal Residual Vector

5.5.3 Checking the Optimal Residual Vector Because the whole manylus in this sector is based on the measurement data that we use for the estimation, we will not be able to completely assume the question of model validation, analysi of ar model is able to make valid-predictions for new simulators. For model validation, we waited the automic ref of measurement data that were not involved in the estimation procedure, for example a new experiment that is performed after the estimation procedure is finishipsed, or a previously concluded experiment that was kept secret during the parameter estimation interver, what we can do with the estimation gate and the single experiment that we use for parameter estimation is to look at the residual values $R_i(\theta^n)$ for i = 1, ..., N. If we plot them as a function of the system of the biological of the residual values $R_i(\theta^n)$. If the histogram tooks that the histogram of a zero mean Gassian with obsolution of the residual values $R_i(\theta^n)$. If the histogram tooks that the through the system core pair of the solution of the residual values $R_i(\theta^n)$. If the histogram took has the the histogram of a zero mean distance of the same that one of the value of the same that a superior that the system core noise model is a distance of the residual values $R_i(\theta^n)$. If the histogram took has the the histogram of a negative that the same that the parameter estimation is the parameter estimation of the residual values $R_i(\theta^n)$ is the first parameter due to be the the the same due to be the the the same due to be same due to be the the the same due to be the the the same due to be due to be the the the same due to be same due to be the the the same due to be the the due to be the the same due to be the the same due to be due to be the the the same due to be due to be the the the same due to be due to be the the the same due to be due to be the the the due due to be the the the due to be the the due to be the the the the due to be the the the due to be the the the due t

How certain is our estimate?

When we formulate and solve a nonlinear least squares problem, we need to use a numerical optimization routing to find the maximum likelihood estimate. In order to do this, we first scale the vector of model-measurement mismath residuals $y = M(\theta)$ by using a - usually diagonal - guess Σ_{s} of the covariance matrix of the noise, it

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM

order to obtain the scaled residual vector $R(\theta) := \Sigma_i^{-\frac{1}{2}} (M(\theta) - y)$ such that the maximum likelihood estimate $\hat{\theta} = \theta^*$ is obtained by the solution of the optimization problem

 $\theta^* = \arg \min_{\theta} \frac{\frac{1}{2} \|R(\theta)\|_2^2}{-f^{(\theta)}}$

(5.30)

Note that the residual function R maps from \mathbb{R}^d to \mathbb{R}^N , and that most solution algorithms require the user to pass this function $R(\theta)$ – and not the objective function $f(\theta) = \frac{1}{2} \|R(\theta)\|_2^2$ – to the solver. We want to answer three questions in this section:

How do we solve the nonlinear least squares optimization problem (5.30) in practice?
 How can we obtain an estimate of the covariance matrix of the parameter estimate?

How can we obtain an examine or the communic matrix of the parameter estimated
 How can we assess if the modelling assumptions, in particular on the noise, were correct?

We will answer the three questions in the following three subsections.

5.5.1 The Gauss-Newton Algorithm

In practice, noticitor less squares problems are solved with specialized notificat optimization routines like MAT LAB's largenoilin, which expect the user to provide an initial gases for the parameter $\theta -$ which we call θ_{ijk} in this section – and a partient to the functions $P_i(X^{-1} + X^{-1})$. Note studied applicitum, comparative to find call minima. Starting at the initial gases θ_{ijk} , they generate a sequence of iterates that we call θ_{ijk} $\theta_{ijk} \theta_{ijk}$,...,Note that each θ_{ijk} is a section in the specific space of the stars parameters m_{ijk} in the fact in order to distinguals fitting the initial gases θ_{ijk} , they generate a sequence of iterates that we call $\theta_{ijk} \theta_{ijk} \theta_{ijk}$,...,Note that suifies all each their dot processing optimality condition, i.e., as a point that studies relies as a point θ that suifies all each their dot processing optimality condition, i.e., as a point structure the procession of the structure structures r_{ijk} (the distinct structure structure) and the structure structure structure structures the structure structure structure structure structure structure structure structures the structure structure

Idea: Because we know very well how to solve linear least squares problems and because all nonlinear functions on locally be approximated by their first order Taylor series, a straightforward line, would be travely a time least three conversion (in any volcation games θ_{ijk} we have that $M(\theta_i) = H(\theta_{ijk}) + \frac{H_{ijk}}{2\theta_i} |\theta_{ijk}\rangle = \theta_{ijk} + O(||\theta - \theta_{ijk}|)$, and if we use the first order Taylor series to formulate a linear least games problem to order to find β_{ijk} , resp.

 $\begin{array}{ll} \theta_{[k+1]} &= \arg \min_{\theta} \frac{1}{2} \left\| R(\theta_{[k]}) + \frac{\partial R}{\partial \theta_{[k]}} (\theta_{[k]}) (\theta - \theta_{[k]}) \right\|_{2}^{2} & (5.31) \\ &= \arg \min_{\theta} \frac{1}{2} \left\| - - (\theta_{[k]}) \theta_{[k]} + R(\theta_{[k]}) + J(\theta_{[k]}) \theta \right\|_{2}^{2} & (5.32) \\ &= (J(\theta_{[k]})^{2} J(\theta_{[k]})^{-1} J(\theta_{[k]})^{-1} (J(\theta_{[k]}) \theta_{[k]} - R(\theta_{[k]})) & (5.33) \end{array}$

 $= (J(\theta_{[k]})^\top J(\theta_{[k]}))^{-1} J(\theta_{[k]})^{-1} (J(\theta_{[k]}), \theta_{[k]} - R(\theta_{[k]}))$ (5.33) $= \theta_{[k]} - (J(\theta_{[k]}))^\top J(\theta_{[k]}))^{-1} J(\theta_{[k]})^{-1} R(\theta_{[k]})$ (5.34) $= \theta_{[k]} - J(\theta_{[k]})^{-} R(\theta_{[k]})$ (5.35)

Note that the iteration above is only well defined if the Jacobian matrix $J_0(\theta_{00})$ is of fail rank, but that is prerestical implementations, small algorithm molfications ensures that teach thransion but defined. With the above expression, we have already defined the basic Gauss-Newton algorithm. Once can show that – if a converges – facuss-Newton algorithm converges linearly to a satisotary group of with $D_1^{(0)}(0) = 0$. But a proof of this small is byoind one interest here.

52 CHAPTER 5. MAXIMUM LIKEL BHOOD AND BAYESIAN ESTIMATION f, which are given by $f(\theta) = \frac{1}{2} ||R(\theta)|_{2}^{2} = \frac{1}{2} \sum_{k=1}^{N} R_{k}(\theta)^{2} \qquad (5.36)$ $\nabla f(\theta) = \sum_{k=1}^{N} \nabla R_{k}(\theta) R_{k}(\theta) - J(\theta)^{-} R(\theta) \qquad (5.37)$ $\nabla^{2} f(\theta) = J(\theta)^{-} J(\theta) + \sum_{k=1}^{N} \nabla^{2} R_{k}(\theta) R_{k}(\theta) \qquad (5.36)$ Using some of the above expressions, the letterians of the Gauss-Newton algorithm can be written as

 $\theta_{[k+1]} = \theta_{[k]} - B_{GN}(\theta_{[k]})^{-1}\nabla f(\theta_{[k]})$

It can be seen, as expected from a origination algorithm, but they point with $\nabla_i^*(\theta_{ijk}) = 0$. But the interest matrix $B_{ijk}(\theta_{ijk})^{-1}$ in from of the gradient could also be hower differently. It loss would choose the interest of the assist matrix, $\nabla_i^*(\theta_{ijk})^{-1}$, one would choose the two sources of the second second second methods, which comparisate for many Gause-Newton methods. The matrix $D_{ijk}(\theta_{ijk})$ is called the different choices of Hessian approximations for matrix $B_{ijkk}(\theta_{ijk})^{-1}$, and would choose in the so-called Newton worked, different choices of Hessian approximations the family of allows. Newton methods. The matrix $D_{ijkk}(\theta_{ijk})$ is called the dimension in the state of the origin of the source of the source of the large comparison of the source of the

5.5.2 Estimating the Covariance Matrix and Extracting its Relevant Entries

The usine way to obtain a mappi estimate of the correlations matrix Σ_0 of the parameter estimate T^0 would be isomer that the biseculation of R at the obtain is the correct model, and that all the catalical assumptions with matrix Σ_0 . Following the formulation of the fraction R were correct, i.e., that we indeed had Gaussian noise with covariant matrix Σ_0 . Following the interfaced system analysis, we could then fixely use $R(S_0)^{-1} \to 1$ parameters Z_0 . Following the interfaced system $R(\theta) = \Sigma_0^{-1} (R(\theta) - \mu)$, correstance matrix $\Sigma_0 \to 10^{10}$ model is the state of the single system components in our observation of the order of one. For the variant of the single system interfaced matrix $Z_0 \to 10^{10}$ model are state as a parameter of the single system interfaced matrix $Z_0 \to 10^{10}$ model are $Z_0 \to 10^{10}$ model and $Z_0 \to 10^{10}$ model are $Z_0 \to 10^{10}$ model. The first $Z_0 \to 10^{10}$ model matrix $Z_0 \to 10^{10}$ model model is a state of the single system interiming the restrict $Z_0 \to 10^{10}$ model model are $Z_0 \to 10^{10}$ model model and $Z_0 \to 10^{10}$ model model model and $Z_0 \to 10^{10}$ model mod

$$\begin{split} \Sigma_{ij} &:= \frac{|R(\theta^*)|_{ij}^2}{N-d} \left(I(\theta^*)^{-1}(\theta^*) \right)^{-1} \end{split}$$
 (one wants to express the result of the parameter estimation, one often only uses the diagonal entries from this atrix, which contain, for $i=1,\ldots,d$, the valances $\hat{\sigma}_i^2$ (of the respective parameter components $\hat{\theta}_n$, as seen in the obliving detailed matrix expression: $\Sigma_{ij} = \begin{bmatrix} \sigma_i^2 & \bullet & \bullet \\ \bullet & \bullet & \ddots \\ \bullet & \bullet & \bullet \end{bmatrix}$

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SOUARES PROBLEM

Taking the square root of the variances yields the standard deviations, such that the final result of the whole arameter estimation procedure could be presented by only 2d numbers in the form $\theta_i = \theta_i^a \pm \sqrt{\sigma_i^a}$, for i = 1,...,d

$v_i = v_i + \sqrt{v_i}$, in 5.5.3 Checking the Optimal Residual Vector

Sciences the whole analysis in this section is based on the measurement data that we use for the estimation, we will not be able to completely answer the question of model validation, namely of our model is able to make valid expected on the section of the estimation of the section of the s

to look at the residual values $R_i(\theta^m)$ for i = 1, ..., N. If we plot them as a function of i, they should look its a segment of moder mankers. In order to check this in more details, one spiculity creates and plots a sinsignan of the residual values $R_i(\theta^m)$. If the bisogram looks like the bisogram of a zero mean Gaussia with its variance, the model assumptions on the system and moles are likely be occurred. If not, some grant of the modelling assumptions was probably wrong. One should then think hard and change the system mode in results, but on catterer model.

Estimate:

$$\Sigma_{\hat{ heta}} := rac{\|R(heta^*)\|_2^2}{N-d} \left(J(heta^*)^{ op}J(heta^*)
ight)^{-1}$$

- Works for linear systems and Gaussian distributions
- May work for nonlinear systems
 - No guarantees!

When we formulate and solve a nonlinear least squares problem, we need to use a numerical optimization routine to find the maximum likelihood estimate. In order to do this, we first scale the vector of model-measurement-mismath residuals $y = M(\theta)$ by using a – usually diagonal – geose S, of the covariance matrix of the noise, in

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM

order to obtain the scaled residual vector $R(\theta) := \Sigma_c^{-\frac{1}{2}}(M(\theta) - y)$ such that the maximum likelihood estimate $\hat{\theta} = \theta^*$ is obtained by the solution of the optimization problem

 $\theta^* = \arg \min_{\theta} \frac{\frac{1}{2} \|R(\theta)\|_2^2}{-f(\theta)}$

\$1

(5.30)

Note that the residual function R maps from \mathbb{R}^d to \mathbb{R}^N , and that most solution algorithms require the user to pass this function $R(\theta) = and$ not the objective function $f(\theta) = \frac{1}{2} \|R(\theta)\|_{1}^2$ – to the solver. We want to answer three questions in this section:

· How do we solve the nonlinear least squares optimization problem (5.30) in practice?

· How can we obtain an estimate of the covariance matrix of the parameter estimate? · How can we assess if the modelling assumptions, in particular on the noise, were correct?

We will answer the three questions in the following three subsections.

5.5.1 The Gauss-Newton Algorithm

2.5.1 I the UBBN-VEWON Approximate Is practice, online lesis spaces problems serviced with specialized notificate optimization nontines like MAT-LAPS 1 support 11, which expect the user to provide an initial guess for the parameter θ – which we call θ₁, in this section – and a pointer to the functions R R² → R². Note subthe algorithms can only guestrate to final location for R R² → R². Note that the section of an distribution of the section of an distribution of R R² → R². Note that the section of a pointer to the function R R² → R². Note that exceed θ₁ is a total field section of the section of a section of the section of a section of the section of a section of the section o

Because we know very well how to solve linear least squares problems and because all nonlinear functions by be approximated by their first order Taylor series, a straightforward idea would be to solve a finance least strength makes on the interaction at a solution gene θ_{ii} in order to do that a better value for gene λ_{ii} netreity, for any solution guess θ_{ii} , we have that $R(\theta) = R(\theta_i) + \frac{2\pi}{2\theta}(\theta_{ii})(\theta - \theta_{iii}) + O((\theta - \theta_{iii})))$.

$\theta_{[k+1]} =$	$\arg \min_{\theta} \frac{1}{2} \left\ R(\theta_{[k]}) + \underbrace{\frac{\partial R}{\partial \theta}(\theta_{[k]})(\theta - \theta_{[k]})}_{- \cup (\theta_{[k]})} \right\ _{2}^{2}$	(5.31)
=	$\arg \min_{\theta} \frac{1}{2} \left\ -J(\theta_{[k]}) \theta_{[k]} + R(\theta_{[k]}) + J(\theta_{[k]}) \theta \right\ _{2}^{2}$	(5.32)
-	$(J(\theta_{[k]})^{\top}J(\theta_{[k]}))^{-1}J(\theta_{[k]})^{\top}$ $(J(\theta_{[k]})$ $\theta_{[k]} - R(\theta_{[k]}))$	(5.33)

-	$\theta_{[k]} - (J(\theta_{[k]})^\top J(\theta_{[k]}))^{-1} J(\theta_{[k]})^\top R(\theta_{[k]})$	(5.34)
-	$\theta_{[k]} - J(\theta_{[k]})^+ R(\theta_{[k]})$	(5.35)
	ly well defined if the Jacobian matrix $J(\theta_{ k })$ is of full rank	

small algorithm modifications ensure that each iteration is well defined. With the above eady defined the basic Gauss-Newton algorithm. One can show that - if it converges - the or converges linearly to a stationary point ∂^2 with $\partial^2/(\partial^2) = 0$, but a proof of this result is or o understand the algorithm a bit better and to see at least why the algorithm does not move point, it is useful to look at explicit expressions for the derivatives of the objective function

52	CHAPTER 5. MAXIMUM LIKELIHOOD AND BAY	SIAN ESTIMATION
f, which are given by		
	$f(\theta) = \frac{1}{2} R(\theta) _2^2 = \frac{1}{2} \sum_{i=1}^{N} R_i(\theta)^2$	(5.36)

$\nabla f(\theta)$	-	$\sum_{i=1}^{N} \nabla R_{i}(\theta) R_{i}(\theta) = J(\theta)^{\top} R(\theta)$	(5.37)
$\nabla^2 f(\theta)$	-	$\underbrace{J(\theta)^{\top}J(\theta)}_{=R_{i}(\theta)} + \sum_{i=1}^{N} \nabla^{2}R_{i}(\theta)R_{i}(\theta)$	(5.38)

Jsing some of the above expressions, the iterations of the Gauss-Newton algorithm can be written as

 $\theta_{[k+1]} = \theta_{[k]} - B_{GN}(\theta_{[k]})^{-1} \nabla f(\theta_{[k]})$

It can be seen, as expected from a certification algorithm $(\nabla G(m_{0}), \nabla T(\mathcal{B}_{0}))$ It can be seen, as expected from as explanations algorithm, but the algorithm would not move away from a stationary point with $\nabla(\mathcal{B}_{0,1}) = 0$. But the inverted matrix $B_{0,1}((k_{0,1})^{-1})$ for bot of the gradient could also be hower differently. It was would choose the inverse of the exact Heasin matrix $\mathcal{B}_{0,1}((k_{0,1})^{-1})$ can would chain the two scale of the solid lenses are predicted and the solid lense of the solid lenses approximation in fract comparison of the actual function. The actual function, the cause of the first control relation specific lenses of the solid lense of the solid lenses of the solid

5.5.2 Estimating the Covariance Matrix and Extracting its Relevant Entries

5.5.2 Holimating the Covariance Natirx and Extracting its Keirvant Entries The exist says to obtain a rough estimate of the covariance matrix S_0 of the parameter estimate $\beta^{(1)}$ would be to assume that the linearization of R at the solution is the correct model, and that all the statistical assumptions we made in the formation of the function N were coverie. (i.e., that we indeed had Gaussian noise with covariance matrix J_0 . Following the linear least squares analysis, we could then directly use $B_{(5)}(\theta^{(1)})^{-1}$ is presented rotations, the spring squared residual value is expected to be in the order of N. More prescript, due the fact that we have a dimensional vector fitting the data and minimizing the residual, we receively, due the fact that we have a dimensional vector fitting the data and minimizing the residual. We receively, due the fact that we have a dimensional vector fitting the data and minimizing the residual. We receively due the fact that we have a dimensional vector fitting the data and minimizing the residual. We receively due the fact that we have a dimensional vector fitting the data and minimizing the residual. We receively due the fact that we have a dimensional vector fitting the data and minimizing the residual. We receively due the fact that we have a dimensional vector fitting the data and minimizing the residual.

 $\Sigma_{\hat{\theta}} := \frac{\|R(\theta^*)\|_2^2}{N-d} \left(J(\theta^*)^\top J(\theta^*)\right)^{-1}$ If one wants to express the result of the parameter estimation, one often only uses the diagonal entries from this matrix, which contain, for i = 1, ..., d, the variances σ_i^2 of the respective parameter components θ_{ii} as seen in the following detailed matrix expression:



5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM 53

Taking the square root of the variances yields the standard deviations, such that the final result of the whole parameter estimation procedure could be presented by only 2d numbers in the form $\theta_i = \theta_i^* \pm \sqrt{\sigma_i^2}$, for i = 1, ..., d

5.5.3 Checking the Optimal Residual Vector

5.5.3 Checking the Optimal Residual Vector Because the whole marks in this section is based on the measurement data that we use for the estimation, we will not the able to completely answer the question of model validation, numbly if nor model is able to make value predictions for new simulators. For model validation, we want the quarkness ref or measurement data that were not involved in the estimation procedure, for example a new coperiment that is performed after the estimation procedure is finished, or a previously control of questions that was keep secret diminity the parameter estimation, it to solve at the residual value $R_i(\theta^n)$ for i = 1, ..., N. If we plot them as a function of it, they should look the a sequence of random numbers, but whole to check theirs in more dealing, one typically creases and plots a like a sequence of random numbers, but whole to hole the time the distribution of the systemat con-ductions, the model assumptions on the system and nuise are likely be corner. If not, some part of the modeling assumptions was probably wavego. One should dotted that has a different model and restart the parameter estimation procedure, based on the same data, but on a different model.

 $\Sigma_{\hat{\theta}} = \begin{bmatrix} \sigma_1^2 & * & * & * \\ * & \sigma_2^2 & * & * \\ & & \ddots & * \\ * & * & \ddots & * \\ \vdots & \vdots & \vdots & * & \sigma_2^2 \end{bmatrix}$ σ^2

 $heta_i = heta_i^* \pm \sqrt{\sigma_i^2}$

In our case:

$$a = 0 \pm 0.07$$

```
sigma_est = sqrt( (R(a)'*R(a))/(length(y)-1) * inv(J(a)'*J(a)) )
```

% We can generate a "measurement" % by using M(0) as the "ground truth" and adding % random noise to it. % (Uncomment the line below to expermient with % estimating a from a different measurement.) % y = M(0) + S * randn(size(M(0)))

% ... and re-run the simulation

MOAR PARAMETERS!



- Maybe there's an offset in all the measurements?
- Let's introduce another parameter!

New ML problem



- $\sin(a) + b + \epsilon_1 = -0.1$ $\sin(a + 30^\circ) + b + \epsilon_2 = 0.6$ $\sin(a + 60^\circ) + b + \epsilon_3 = 0.9$
 - All
 e are independent, and drawn from a normal distribution with zero mean and standard deviation 0.5

```
M = @(a, b) \sin(a + pi/6 * [0 1 2]') + b
```

```
% Note, if you want to see this script produce useful estimates,
% then the data would have to be drawn from a distribution with
% a lot smaller variance. (That is: C should be smaller.)
C = diag(0.5^2 * [1 1 1])
S = sqrtm(C)
y = [-0.1, 0.6, 0.9]'
y = M(0, 0) + S * randn(size(M(0, 0)))
R = @(a, b) inv(S) * (M(a, b) - y)
p = Q(a, b) exp(-0.5*sum(R(a, b).^2))
as = -pi:0.01:pi;
bs = -3:0.01:3;
ps = zeros(length(as), length(bs));
for i=1:length(as)
  for j=1:length(bs)
   ps(i,j) = p(as(i), bs(j));
  end
end
figure(1)
% Plot integral of p over all b values, as function of a
plot(as, sum(ps, 2))
figure(2)
% Plot integral of p over all a values, as function of b
plot(bs, sum(ps, 1))
figure(3)
% Plot p as function of a and b
contour(bs, as, ps)
[pmax, ix] = max(sum(ps, 2))
a = as(ix)
[pmax, ix] = max(sum(ps, 1))
b = bs(ix)
t = [a, b]'
J = @(t) inv(S) * [cos(a + pi/6 .* [0 1 2]'), [1 1 1]']
R2 = @(t) R(t(1), t(2))
t = t - J(t) \setminus R2(t)
C est = (R2(t)'*R2(t))/(length(y)-2) * inv(J(t)'*J(t))
sigma_est = sqrt(diag(C_est))
```

When we formulate and solve a nonlinear least squares problem, we need to use a numerical optimization routine to find the maximum likelihood estimate. In order to do this, we first scale the vector of model-measurementmismath residuals $y = M(\theta)$ by using a – usually diagonal – geose S, of the covariance matrix of the noise, in

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM

order to obtain the scaled residual vector $R(\theta) := \Sigma_r^{-\frac{1}{2}}(M(\theta) - y)$ such that the maximum likelihood estimate $\hat{\theta} = \theta^*$ is obtained by the solution of the optimization problem

 $\theta^* = \arg \min_{\theta} \frac{\frac{1}{2} \|R(\theta)\|_2^2}{-f(\theta)}$

\$1

(5.30)

Note that the residual function R maps from \mathbb{R}^d to \mathbb{R}^N , and that most solution algorithms require the user to pass this function $R(\theta)$ – and not the objective function $f(\theta) = \frac{1}{2} \|R(\theta)\|_2^2$ – to the solver. We want to answer three questions in this section:

How do we solve the nonlinear least squares optimization problem (5.30) in practice?

How can we obtain an estimate of the covariance matrix of the parameter estimate?
 How can we assess if the modelling assumptions, in particular on the noise, were correct?

We will answer the three questions in the following three subsections.

5.5.1 The Gauss-Newton Algorithm

Solid The Control region region maps in the control region maps in the control region region in the control region region the other control region reg

Here: Because we know very well how to solve linear least squares problems and because all moliner functions calculable approaches that the solution of the s

$\theta_{[k+1]} =$	$\arg \min_{\theta} \frac{1}{2} \left\ R(\theta_{[k]}) + \underbrace{\frac{\partial R}{\partial \theta}(\theta_{[k]})(\theta - \theta_{[k]})}_{-J(\theta_{[k]})} \right\ _{2}^{2}$	(5
-	$\arg \min_{\theta} \frac{1}{2} \left\ -J(\theta_{[k]}) \theta_{[k]} + R(\theta_{[k]}) + J(\theta_{[k]}) \theta \right\ _{2}^{2}$	(5
-	$(J(\theta_{[k]})^{\top}J(\theta_{[k]}))^{-1}J(\theta_{[k]})^{\top}(J(\theta_{[k]}) \theta_{[k]} - R(\theta_{[k]}))$	(5

 $= \begin{pmatrix} \psi_{ijk}(r + \psi_{ijk}) - \psi_{ijk}(r) \\ = \theta_{ijk}(-r)(\theta_{ijk})^T J(\theta_{ijk})^{-1} J(\theta_{ijk})^T R(\theta_{ijk}) \\ = \theta_{ijk} - R(\theta_{ijk})^T R(\theta_{ijk}) \\ = \delta_{ijk} - R(\theta_{ijk})^T R(\theta_{ijk}) \\ \text{(5.31)}$ ration above is only well defined if the Jacobian matrix $J(\theta_{ijk})$ is of full rank, but their practions, small algorithm modifications essure that each hermion is well defined. With the above

tical implementations, small algorithm modifications ensure that each institute is well defined. With the above expression, we have estimately defined the have (Gauss-Newton algorithm. One can show that – if a convergention of the strengthm convergence intensity to a stationary point θ^{*} with $\nabla f(\theta^{*}) = 0$, but a proof of this result is beyond on interest be them. However, in order to understand the algorithm a bit thetar and to see at least why the algorithm does not move any prior. In stationary point, it is useful to a station for the derivatives of the depicetive functions.

52	CHAPTER 5. MAXIMUM LIKELIHOOD AND BAY	SIAN ESTIMATION
f, which are given by		
	$f(\theta) = \frac{1}{2} R(\theta) _2^2 = \frac{1}{2} \sum_{i=1}^{N} R_i(\theta)^2$	(5.36)

$\nabla f(\theta)$	-	$\sum_{i=1}^{N} \nabla R_{i}(\theta) R_{i}(\theta) = J(\theta)^{\top} R(\theta)$	(5.37)
$\nabla^2 f(\theta)$	-	$\underbrace{J(\theta)^{\top}J(\theta)}_{=:B_{cw}(\theta)} + \sum_{i=1}^{N} \nabla^{2}R_{i}(\theta)R_{i}(\theta)$	(5.38)

ations of the Gauss-Newton algorithm can be

 $\boldsymbol{\theta}_{[k+1]} = \boldsymbol{\theta}_{[k]} - B_{\mathrm{GN}}(\boldsymbol{\theta}_{[k]})^{-1} \nabla f(\boldsymbol{\theta}_{[k]})$

It can be seen, as expected from an equivalent of the algorithm would not move away from a statienty point with $\nabla f(\theta_{11}) = 0$. But the inverted matrix $R_{12}(\theta_{12})^{-1}$ in front of the gradient could also be chosen differently. To see void chosen the inverse of the exact Resina matrix $T(\theta_{12})^{-1}$ in front of the gradient could also be chosen differently. The would chosen be inverse of the exact Resina matrix $T(\theta_{12})^{-1}$ in the own de chain the total set of the resinate prove that indicates the final set of the resinate provident in the set of the resinate the resinate the set of the resinate the resinates the resi

5.5.2 Estimating the Covariance Matrix and Extracting its Relevant Entries

The exists way to obtain a rough estimate of the covariance matrix Σ_{ij} of the parameter estimate β^{ij} would be as empty the existing obtained on the covariance matrix Σ_{ij} of the parameter estimate β^{ij} would be as empty obtained on the formulation of the fraction R were cover. i.e., that we indeed fuel Gaussian noise with overained matrix Σ_{ij} . Following the limits role as signary assume subsylve, we could then directly use $|R_{ij}(\theta)|^{-1}$ by a generative covariance matrix. Note that, due to the scaling in the expression $R(\theta) = \Sigma_{ij}^{-1} (M(\theta) - \mu_{ij})$, or assumption would be that the scale of the residual vector empowers in so the outper scale. Note that, due to the scaling is not only unless, but not in the due of one. For this reasor, the optimal squared residual value is expected to be in the order of N. More precisely, due fact that we as a distinguishing the vector fitting the duar minimizing the models. We respect ($|P_{ij}|| \le N - 4$, as introduce the scale of the scale scale and minimizing the models. We respect ($|P_{ij}|| \le N - 4$, as introduce the scale of the squared residual ($|P_{ij}|| \le N - 4$). Since the different from N - 4. Because this is expected residual ($|P_{ij}|| \le N - 4$). Because the scale is a scale scal

 $\Sigma_{\phi} := \frac{\|R(\theta^*)\|_2^2}{N-d} \left(J(\theta^*)^\top J(\theta^*)\right)^{-1}$

If one wants to express the result of the parameter estimation, one often only uses the diagonal entries from this matrix, which contain, for i = 1, ..., d, the variances σ_i^2 of the respective parameter components $\hat{\theta}_{ii}$ as seen in the billowing detailed matrix expression:

ion: $\Sigma_{\tilde{g}} = \begin{bmatrix} \sigma_1^2 & \cdot & \cdot \\ \sigma_2^2 & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \sigma_d^2 \end{bmatrix}$

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM 53

king the square root of the variances yields the standard deviations, such that the final result of the whole rameter estimation procedure could be presented by only 2d numbers in the form $\theta_i = \theta_i^* \pm \sqrt{\sigma_i^2}$, for i = 1,...,d

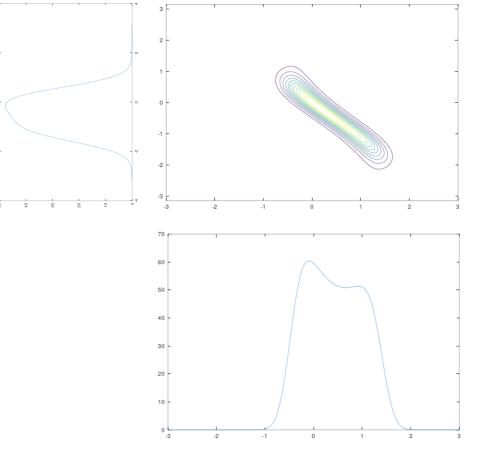
5.5.3 Checking the Optimal Residual Vector

5.5.5 CircCating use Opinian Restructure Vectors in the ensurement data that we use for the estimation, we will not be able to completely assess the aposition of model wildinkon, namely if our model is able to make wild not be able to completely assess the aposition of model wildinkon the second on the second sec

procedure and was just reserved for model validation. Bowever, what we can do with the civiting data of the single experiment that we use for parameter estimation, is to look at the residual values $R_i(P^*)$ for $i = 1, ..., N_i$ the plot them as a fraction of *i*, they should look kinks a sequence of moden numbers. In a order to check this in modellas, one septimity creates and plots a bilitary and the residual values $R_i(P^*)$. If the biasogram looks like the biasogram of a zero meas Gaussian with mixediment assumptions on the system and molass are likely to be correct. In our, some part of the modellas susemptions was probably wrong. One should then think hard and drange the system or noise model and result the material estimation of the size of the si $\Sigma_{\hat{\theta}} = \begin{bmatrix} \sigma_1^2 & * & * & * \\ * & \sigma_2^2 & * & * \\ & & \ddots & \\ * & * & \ddots & * \\ * & * & * & \sigma_d^2 \end{bmatrix}$

 $\theta_i = \theta_i^* \pm \sqrt{\sigma_i^2}$

 $b = -0.13 \pm 0.41$



 $a = -0.14 \pm 0.48$

When we formulate and solve a nonlinear least squares problem, we need to use a numerical optimization routin to find the maximum likelihood estimate. In order to do this, we first scale the vector of model-measurement mismath residuals $y = M(\theta)$ by using a – usually diagonal – grees Σ_{o} of the covariance matrix of the neise, i.

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SOUARES PROBLEM

aled residual vector $R(\theta) := \sum_{e}^{-\frac{1}{2}} (M(\theta) - y)$ such that the maximum likelihood estimat

(5.30)

Note that the residual function R maps from \mathbb{R}^d to \mathbb{R}^N , and that most solution algorithms require the user to pass his function $R(\theta) = \operatorname{and}$ not the objective function $f(\theta) = \frac{1}{2} |R(\theta)|_2^2 - to the solver. We want to answer three$ metrics in this section:

· How do we solve the nonlinear least squares optimization problem (5.30) in practice? · How can we obtain an estimate of the covariance matrix of the parameter estimate?

· How can we assess if the modelling assumptions, in particular on the noise, were correct?

We will answer the three questions in the following three subsections.

5.5.1 The Gauss-Newton conservation in the specialized nonlinear least squares problems are solved with specialized nonlinear LAB's 1 square 1, which expect the user to provide an initial gauss for the provide a nonline to the function $R: \mathbb{R}^d \to \mathbb{R}^n$. Note available algorithm of the neutron of the

 $\theta_{[k+1]} = \arg \min_{\theta} \frac{1}{2} \left\| R(\theta_{[k]}) + \frac{\partial R}{\partial \theta}(\theta_{[k]})(\theta - \theta_{[k]}) \right\|_{2}^{2}$ $= \arg \min \frac{1}{\pi} \left\| -J(\theta_{(k)}) \theta_{(k)} + R(\theta_{(k)}) + J(\theta_{(k)}) \theta \right\|^2$ (5.32) $= (J(\theta_{[k]})^{\top}J(\theta_{[k]}))^{-1}J(\theta_{[k]})^{\top} (J(\theta_{[k]}) \theta_{[k]} - R(\theta_{[k]}))$ $\theta_{[k]} = (J(\theta_{[k]})^\top J(\theta_{[k]}))^{-1} J(\theta_{[k]})^\top R(\theta_{[k]})$ (5.34)

 $= \theta_{(k)} - J(\theta_{(k)})^+ R(\theta_{(k)})$ (5.35) ove is only well defined if the Jacobia tand the algorithm a bit better and to see at least why the al-

CHAPTER 5. MAXIMUM LIKELIHOOD AND BAYESIAN ESTIMATION 52 f, which are given by $f(\theta) = \frac{1}{2} ||R(\theta)||_2^2 = \frac{1}{2} \sum_{i=1}^{N} R_i(\theta)^2$ (5.36) $\nabla f(\theta) = \sum_{i=1}^{N} \nabla R_i(\theta) R_i(\theta) = J(\theta)^\top R(\theta)$ (5.37) $\nabla^2 f(\theta) = \underbrace{J(\theta)^\top J(\theta)}_{i=1} + \sum_{i=1}^N \nabla^2 R_i(\theta) R_i(\theta)$ (5.38) s, the iterations of the Gauss-Newton ala $\theta_{[k+1]} = \theta_{[k]} - B_{GN}(\theta_{[k]})^{-1}\nabla f(\theta_{[k]})$ If from an optimization algorithm, that the algorith $\theta_{|k|} = 0$. But the inverted matrix $B_{cros}(\theta_{|k|})^{-1}$ in t

ngly, it is useful to also obtain from the algorith sian $B_{GN}(\theta^*)^{-1}$, because it can serve as approxi-5.5.2 Estimating the Covariance Matrix and Extracting its Relevant Entries The easiest way to obtain a rough estimate of the covariance matrix Σ_{β} of the parameter estimate θ^* wos assume that the linearization of R at the solution is the correct model, and that all the statistical assump and in the formulation of the function R were correct, i.e., that we indeed had Gaussian noise with co ote that, due to the scaling in the expression $R(\theta) = \Sigma$

size of the squared residual $||R(\theta^*)||_2^2$ can be di asoning of Section 4.7, and in practice use the $\Sigma_{\hat{\theta}} := \frac{\|R(\theta^*)\|_2^2}{N-d} (J(\theta^*)^\top J(\theta^*))^{-1}$ e result of the parameter estimation, one often only uses the diagonal entries from this i = 1, ..., d, the variances σ_i^2 of the respective parameter components $\hat{\theta}_{i_i}$ as seen in the argumentous

 $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}} = \begin{bmatrix} \sigma_1^2 & \bullet & \bullet & \bullet \\ \bullet & \sigma_2^2 & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \sigma_d^2 \end{bmatrix}$

5.5. PRACTICAL SOLUTION OF THE NONLINEAR LEAST SQUARES PROBLEM

uare root of the variances yields the standard deviations, such that the final result of the whole mation procedure could be presented by only 2*d* numbers in the form $\theta_i = \theta_i^* \pm \sqrt{\sigma_i^2}$, for i = 1, ..., d

5.5.3 Checking the Optimal Residual Vector

use the whole analysis in this section is based on the measuren ot be able to completely answer the question of model validation tions for new situations. For model validation, we would nee

Looking at the residuals

>> R(theta(1), theta(2))ans =

0.46141 -0.655790.19438

