Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Fundamental Fault Detection Limitations in Linear Non-Gaussian Systems

Gustaf Hendeby
Division of Automatic Control
Department of Electrical Engineering
Linköpings universitet, SE-581 83 Linköping, SWEDEN
hendeby@isy.liu.se

Fredrik Gustafsson
Division of Automatic Control
Department of Electrical Engineering
Linköpings universitet, SE-581 83 Linköping, SWEDEN
fredrik@isy.liu.se

***Abstract*— Sophisticated fault detection (FD) algorithms often include nonlinear mappings of observed data to fault decisions, and simulation studies are used to support the methods. Objective statistically supported performance analysis of FD algorithms is only possible for some special cases, including linear Gaussian models. The goal here is to derive general statistical performance bounds for any FD algorithm, given a non-linear non-Gaussian model of the system. Recent advances in numerical algorithms for nonlinear filtering indicate that such bounds in many practical cases are attainable. This paper focuses on linear non-Gaussian models. A couple of different fault detection setups based on parity space and Kalman filter approaches are considered, where the fault enters a computable residual linearly. For this class of systems, fault detection can be based on the best linear unbiased estimate (BLUE) of the fault vector. Alternatively, a nonlinear filter can potentially compute the maximum likelihood (ML) state estimate, whose performance is bounded by the Cramér-Rao lower bound (CRLB). The contribution in this paper is general expressions for the CRLB for this class of systems, interpreted in terms of fault detectability. The analysis is exemplified for a case with measurements affected by outliers.**

I. Introduction

In many practical applications it is vital to monitor a parameter which can undergo rapid changes. This is often referred to as *change detection* or *fault detection*. The fault detection problems considered in this contribution can be restated as detecting a nonzero parameter vector in the linear regression model

$$\mathbb{R}_t = H_\theta \theta + H_v \mathbb{V}_t, \tag{1}$$

where $\mathbb{R}_t$ is a residual computed by the fault detection algorithm, $\mathbb{V}_t$ is a stochastic noise term with known *probability density function* (PDF), noise coloring H_v, and H_θ is a regression matrix that depends on the signal model. The hypothesis test for fault detection is stated as

$$\begin{cases} \mathcal{H}_0: & \theta = 0 \\ \mathcal{H}_1: & \theta \neq 0 \end{cases}, \tag{2}$$

where $\mathcal{H}_0$ represent a fault free situation, and $\mathcal{H}_1$ that a fault (change) is present. Section II describes how fault detection in general linear state-space models with additive faults can be reformulated according to (1). The derivation covers both parity space approaches, where state and disturbance are eliminated by projection, as well as Kalman and nonlinear filtering approaches, where the state is estimated.

The general principle for hypothesis testing [1, 2] involves, explicitly or implicitly, estimation of θ. One simple fault detection approach not necessarily based on stochastic theory, is to compute the *least squares estimate* (LSE) of θ and check if it significantly differs from 0. The *weighted* LSE provides the *best linear unbiased estimate* (BLUE) of θ, which enables a test that is more efficient compared to the *non-weighted* LSE in case of colored noise or changing variance. The BLUE compensates fully for second order properties of the noise, but no higher order moments. The *minimum variance estimator* (MVE) and *maximum likelihood estimator* (MLE) are generally nonlinear functions of data, and usually no closed form solution exists. However, recent advances in numerical methods for nonlinear filtering such as the *particle filter* (PF) [3–5] enable on-line MVE and MLE, if sufficient computation time and memory are available. Asymptotically, the MLE attains the *Cramér-Rao lower bound* (CRLB) [6, 7], and the PF obtains, or comes close to, the limit. The performance bound for fault detection discussed in this contribution is based on the CRLB, which is analytically computable for the considered class of systems.

The fundamental question is how much better fault detection performance can be obtained by using a nonlinear MVE or MLE compared to the BLUE. A first answer to this question was presented in [8], and later elaborated on in [9], where estimation and detection in colored non-Gaussian autoregressive (AR) processes are treated. As part of this the *intrinsic accuracy* (IA, see Section II-A for a formal definition) of the PDF of $\mathbb{V}_t$ was used. The basic result is that for a given probability of false alarm, the upper bound in detection performance increases monotonously with IA. The worst case performance is achieved for a Gaussian PDF, in which case the BLUE coincides with MLE and no better performance can be obtained. For all other PDF's the asymptotic MLE outperforms the BLUE.

For linear state-space models, the *Kalman filter* (KF) is the BLUE and in case of non-Gaussian noise the PF may perform better. It is shown in [10] that if the *relative accuracy* (RA, see Section II-A) of either the state noise or the measurement noise is larger than one (*i.e.*, non-Gaussian), then the CRLB decreases and the PF has a potential to outperform the KF. These results are here extended to fault detection, where the actual RA of all involved stochastic terms is found to be explicit terms in fault detection performance measures.

0-7803-9568-9/05/$20.00 ©2005 IEEE

After this introduction, Section II defines information, accuracy, and the models used, and corresponding detection statistics are defined in Section III. Section IV present an application of the theory for data with outliers. Conclusions are found in Section V. The Appendices supplement Section II.

II. Fundamentals

This section introduces *Fisher information* (FI), *intrinsic accuracy* (IA), and *relative accuracy* (RA). Furthermore, this section studies different ways to handle measurements over a time window for linear systems, and gives a common description of the measurements.

A. Information and Accuracy

The *Fisher information* (FI) and *relative accuracy* (RA) described in this section are important for the derivations of detection statistics in Section III.

Definition 1. The *Fisher information* (FI) is defined [6], under mild regularity conditions on the distribution of ξ, as

$$\begin{aligned}\mathcal{I}_\xi(\theta) &:= -\mathsf{E}_\xi\left(\Delta_\theta^\theta \log p(\xi|\theta)\right)\\ &= \mathsf{E}_\xi\left(\left(\nabla_\theta \log p(\xi|\theta)\right)\left(\nabla_\theta \log p(\xi|\theta)\right)^T\right)\end{aligned} \tag{3}$$

evaluated for the true parameter $\theta = \theta_0$, with ∇ and Δ defined to be the *Jacobian* and the *Hessian*, respectively, both defined in Appendix I.

The FI is related to any unbiased estimate of $\hat{\theta}(\xi)$ of θ based on measurements of ξ through

$$\mathsf{cov}\left(\hat{\theta}(\xi)\right) \succeq \mathcal{I}_\xi^{-1}(\theta) = P_\theta^{\text{CRLB}},$$

where P_θ^{CRLB} is the well known *Cramér-Rao lower bound* (CRLB) for the covariance of the estimate $\hat{\theta}$, [1, 6], and $A \succeq 0$ denotes that A is a positive semidefinite matrix.

When nothing else is explicitly stated in this paper, the information is taken to be with respect to the mean, μ assumed to be zero, of the distribution in question, and therefore the notation $\mathcal{I}_e = \mathcal{I}_e(\mu)$, with e a stochastic variable, will be used. This quantity is in [1, 8, 11] referred to as the *intrinsic accuracy* (IA) of the PDF for e. It follows that

$$\begin{aligned}\mathcal{I}_e &= -\mathsf{E}_e\left(\Delta_\mu^\mu \log p_e(e-\mu)\big|_{\mu=0}\right)\\ &= -\mathsf{E}_e\left(\Delta_{e-\mu}^{e-\mu} \log p_e(e-\mu)\big|_{\mu=0}\right)\\ &= -\mathsf{E}_e\left(\Delta_e^e \log p_e(e)\right).\end{aligned}$$

Theorem 1. *For the intrinsic accuracy and covariance of the stochastic variable e the semidefinite inequality*

$$\mathsf{cov}(e) \succeq \mathcal{I}_e^{-1},$$

holds with equality if and only if e has a Gaussian distribution.

Proof. See [12]. □

In this respect the Gaussian distribution is a worst case distribution. Of all distributions with the same covariance the Gaussian is the one with the least information about its mean. All other distributions have larger IA.

Definition 2. If a scalar Ψ_e exists such that $\mathsf{cov}(e) = \Psi_e \mathcal{I}_e^{-1}$, then denote Ψ_e *relative accuracy* (RA) for the distribution.

It follows from Theorem 1 that, when RA is defined, $\Psi_e \geq 1$, with equality if and only if e is Gaussian. The RA is thus a relative measure of how much useful information there is in the distribution, compared to a Gaussian distribution with the same covariance. Other relevant properties of IA are presented in Appendix II.

B. Models

A general structure for stacked residuals is given by

$$\mathbb{R}_t = H_\theta \theta + H_v \mathbb{V}_t, \tag{4}$$

where θ is a structured, low dimensional, fault parameter, *e.g.*, constructed as in Section II-B.2. The matrices Φ, and H_v are system dependent, and $\mathbb{V}_t$ a noise. In the sequel, H_θ and H_v are assumed thick and with full row rank, and $\mathsf{cov}(\mathbb{V}_t) \succ 0$. It is always possible to choose a parameterization such that the nominal parameter is 0. Furthermore, in the next section only (4) will be considered for detection.

The sequel of this section shows how linear regressions, and two variations of detection formulations of dynamic linear systems all can be made to fit in the general linear residual formulation (4).

1) Linear Regression Model: Consider the linear regressions,

$$y_t = \varphi_t^T \theta + e_t,$$

where φ_t is a regression matrix, θ the system parameter with nominal value θ_0, and measurement y_t. Assuming a nominal model θ_0, the residual becomes

$$r_t = y_t - \varphi_t^T \theta_0 = \varphi_t^T \tilde{\theta} + e_t,$$

where $\tilde{\theta} := \theta - \theta_0$ should be interpreted as the model change (fault). Gathering L measurements over a window in time, the regression can be described as

$$\mathbb{R}_t = \Phi^T \tilde{\theta} + \mathbb{E}_t, \tag{5}$$

where $\mathbb{R}_t = \begin{pmatrix} r_{t-L+1}^T & \cdots & r_t^T \end{pmatrix}^T$, Φ stacked regression matrices, and $\mathbb{E}_t$ stacked measurement noise.

2) State-Space Model: In the more general dynamic linear state-space model the measurements are described by the relations:

$$x_{t+1} = A_t x_t + B_{u,t} u_t + B_{w,t} w_t + B_{f,t} f_t, \tag{6a}$$

$$y_t = C_t x_t + D_{u,t} u_t + e_t + D_{f,t} f_t, \tag{6b}$$

where u_t is considered a known deterministic input signal, w_t process noise, e_t measurement noise, and f_t a deterministic, but unknown fault input.

Gathering L measurements over time yields:

$$\mathbb{Y}_t = \mathcal{O} x_{t-L+1} + H_w \mathbb{W}_t + \mathbb{E}_t + H_u \mathbb{U}_t + H_f \mathbb{F}_t, \tag{7}$$

with $\mathbb{Y}_t$, $\mathbb{U}_t$, $\mathbb{W}_t$, and $\mathbb{F}_t$ stacked version of y_t, u_t, w_t and f_t, respectively. Furthermore, $\mathcal{O}$ is the extended observabibilty matrix,

$$\mathcal{O} = \begin{pmatrix} C^T & (CA)^T & \dots & (CA^{L-1})^T \end{pmatrix}^T,$$

H_w, H_u, and H_f are *Toeplitz* matrices describing how w, u, and f, respectively, enter the measurements,

$$H_\star = \begin{pmatrix} D_\star & 0 & \dots & 0 \\ CB_\star & D_\star & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ CA^{L-2}B_\star & CA^{L-3}B_\star & \dots & D_\star \end{pmatrix},$$

for $\star \in \{w, u, f\}$. All these may be time dependent, but this is left out for notational clarity. Finally, x_{t-L+1} is the initial state of the measurement window. For a more complete description of this way to view the system see *e.g.*, [13, 14].

In order to get a low order parameterization of the fault profile, and a non-ambiguous distinction between fault and noise, assume that the fault profile is a smooth function (rather than noise). That is, let $f_t = F^i m_t$, where F^i defines a certain fault direction, and where m_t is the scalar time-varying magnitude. To further structure the fault, choose basis functions φ_t of smooth functions (*i.e.*, polynomials), to model incipient variations of the magnitude $m_t = \varphi_t^T \theta$, where θ has relatively low dimension ($\dim(\theta) \ll L$). For simplicity, assume an orthonormal basis, (*e.g.*, discrete *Chebyshev polynomials*), such that $\sum_{k=t-L+1}^{t} \varphi_k \varphi_k^T = I$. In that case, the fault energy is preserved so $\|m_t\|_2^2 = \sum_{k=t-L+1}^{t} m_k^2 = \|\theta\|_2^2$. Using this notation,

$$f_t = F^i \varphi_t^T \theta \tag{8}$$

and $\bar{B}_{f,t} := B_{f,t} F^i \varphi_t^T$, $\bar{D}_{f,t} := D_{f,t} F^i \varphi_t^T$, and θ replace $B_{f,t}$, $D_{f,t}$, and f_t, respectively, in (6). Similarly, $\bar{H}_f$, derived from $\bar{B}_{f,t}$ and $\bar{D}_{f,t}$, and θ should replace $H_{f,t}$ and $\mathbb{F}_t$, respectively, in (7). Note that θ is time invariant even though f_t is not, and due to the lower dimension detection is easier.

a) Prior knowledge about x_{t-L+1}: If an estimate of the initial state in (7) is available from data outside the studied window, then the effect of $\hat{x}_{t-L+1}$ and the known input $\mathbb{U}_t$ can then be removed from the measurements,

$$\begin{aligned} \mathbb{R}_t &= \mathbb{Y}_t - H_u \mathbb{U}_t - \mathcal{O}\hat{x}_{t-L+1} \\ &= \mathcal{O}\tilde{x}_{t-L+1} + H_w \mathbb{W}_t + \mathbb{E}_t + H_f \mathbb{F}_t \\ &= \underbrace{\begin{pmatrix} \mathcal{O} & H_w & I \end{pmatrix}}_{H_v} \underbrace{\begin{pmatrix} \tilde{x}_{t-L+1}^T & \mathbb{W}_t^T & \mathbb{E}_t^T \end{pmatrix}^T}_{\mathbb{V}} + \underbrace{\bar{H}_f}_{H_\theta} \theta, \end{aligned} \tag{9}$$

where $\tilde{x}_{t-L+1} := x_{t-L+1} - \hat{x}_{t-L+1}$ is the error in the estimate, that can be considered to be noise. Any unbiased estimate $\hat{x}_{t-L+1}$ will suffice, but as will be shown, detection performance depends on the quality of the estimate.

b) Parity Space: If no estimate of x_{t-L+1} is available, or if it is unfavorable to use such an estimate, the signal space can be completely removed from the measurements instead. In this way, only changes with effects outside the nominal system are detectable. This is referred to as *parity space* or *analytical redundancy* [15, 16]. To achieve this, construct a projection, $\mathcal{P}_{\mathcal{O}}^{\perp}$, from the measurement space into a lower dimensional residual space such that $\mathcal{P}_{\mathcal{O}}^{\perp}$ is perpendicular to $\mathcal{O}$, *i.e.*, $\mathcal{P}_{\mathcal{O}}^{\perp}\mathcal{O} = 0$, and the covariance of the resulting residual is non-singular. This yields the residuals

$$\begin{aligned} \mathbb{R}_t &= \mathcal{P}_{\mathcal{O}}^{\perp}(\mathbb{Y}_t - H_u \mathbb{U}_t) = \mathcal{P}_{\mathcal{O}}^{\perp}(H_w \mathbb{W}_t + \mathbb{E}_t + H_f \mathbb{F}_t) \\ &= \underbrace{\mathcal{P}_{\mathcal{O}}^{\perp} \begin{pmatrix} H_w & I \end{pmatrix}}_{H_v} \underbrace{\begin{pmatrix} \mathbb{W}_t^T & \mathbb{E}_t^T \end{pmatrix}^T}_{\mathbb{V}} + \underbrace{\mathcal{P}_{\mathcal{O}}^{\perp} \bar{H}_f}_{H_\theta} \theta, \end{aligned} \tag{10}$$

where by construction $\mathsf{cov}(\mathbb{R}_t) \succ 0$, *i.e.*, $\mathcal{P}_{\mathcal{O}}^{\perp} \begin{pmatrix} H_w & I \end{pmatrix}$ has full row rank. Methods to construct $\mathcal{P}_{\mathcal{O}}^{\perp}$ can be found in [17].

III. Fault Detection

In this section, a method to detect changes (faults) is derived, and its explicit asymptotic statistics computed for the different models in Section II. The test derived is the *Wald test*. The approach is then motivated by the optimality of *generalized likelihood ratio test* (GLR test) before detection performance is calculated.

A. Wald Test

A natural approach to fault detection in a system such as (4) is to estimate the fault θ and from the estimate decide if it significantly differs from 0 or not. If $\hat{\theta}$ is a MLE of θ it will asymptotically fulfill

$$\hat{\theta} \overset{a}{\sim} \mathcal{N}\Big(\theta, \big(\Phi^T S^{-1}(\mathbf{\Psi})\Phi\big)^{-1}\Big), \tag{11}$$

where $S(\mathbf{\Psi}) = \mathcal{I}_{H_v \mathbb{V}}^{-1}$, for $\mathbf{\Psi}$ containing all involved RA and $\overset{a}{\sim}$ denotes asymptotically distributed. The same result for a BLUE is

$$\hat{\theta} \overset{a}{\sim} \mathcal{N}\Big(\theta, \big(\Phi^T S^{-1}(\mathbf{1})\Phi\big)^{-1}\Big), \tag{12}$$

where $S^{-1}(\mathbf{1}) = \mathsf{cov}(H_v \mathbb{V})$. That is, $S(\mathbf{\Psi})$ is the IA of $H_v \mathbb{V}$. For the different models discussed in Section II-B, explicit expressions can be derived and the effect of non-Gaussian noise becomes apparent.

1) Linear Regression: For the linear regression described in Section II-B.1 there is only one noise involved, e, hence

$$S(\mathbf{\Psi}) = S(\Psi_e) = \Psi_e^{-1} \mathsf{cov}(\mathbb{E}_t). \tag{13}$$

In this case the inverse of the RA factors out, $S(\Psi_e) = \Psi_e^{-1} S(1)$, *i.e.*, the effect of a non-Gaussian noise is a scaling of $S(\mathbf{\Psi})$ with Ψ_e^{-1} for linear regressions.

2) State-Space Model:

a) Estimated x_{t-L+1}: For a state-space model with estimated x_{t-L+1} there are three stochastic components involved: the estimation error $\tilde{x}_{t-L+1}$, w, and e. In terms of the RA for those,

$$\begin{aligned} S(\mathbf{\Psi}) = S(\Psi_{\tilde{x}_0}, \Psi_w, \Psi_e) &= \Psi_{\tilde{x}_0}^{-1} \mathcal{O}\, \mathsf{cov}(\tilde{x}_{t-L+1}^{\text{BLUE}}) \mathcal{O}^T \\ &+ \Psi_w^{-1} H_w \mathsf{cov}(\mathbb{W}_t) H_w^T + \Psi_e^{-1} \mathsf{cov}(\mathbb{E}_t). \end{aligned} \tag{14}$$

It is shown in [10], that if $\Psi = \Psi_e = \Psi_w$ then $\Psi_{\tilde{x}_{t-L+1}} = \Psi$. Hence, for the special case that the RA is the same for process noise and measurement noise, the inverse of RA factors out as $S(\Psi, \Psi, \Psi) = \Psi^{-1} S(1, 1, 1)$.

b) Parity Space: In the parity space setting the signal part of the model is removed using a projection leaving only Ψ_w and Ψ_e as interesting quantities, hence

$$S(\mathbf{\Psi}) = S(\Psi_w, \Psi_e) = \Psi_w^{-1}\mathcal{P}_{\mathcal{O}}^{\perp}H_w \,\mathsf{cov}(\mathbb{W}_t)H_w^T\mathcal{P}_{\mathcal{O}}^{\perp T} + \Psi_e^{-1}\mathcal{P}_{\mathcal{O}}^{\perp}\,\mathsf{cov}(\mathbb{E}_t)\mathcal{P}_{\mathcal{O}}^{\perp T}. \quad (15)$$

The inverse RA factors out if $\Psi = \Psi_w = \Psi_e$, $S(\Psi, \Psi) = \Psi^{-1}S(1,1)$.

Note that a sufficient condition for $S(\mathbf{\Psi})$ in (13)–(15) to be nonsingular is that the covariance of the measurement noise is positive definite, which in most cases is a natural assumption about the underlying system.

If the estimate $\hat{\theta}$ is normalized, to get unit covariance, an estimate denoted, $\hat{\theta}^{\mathrm{N}}$, is produced with the statistics

$$\hat{\theta}^{\mathrm{N}} = T^{-\frac{1}{2}}(\mathbf{\Psi})\hat{\theta} \overset{a}{\sim} \mathcal{N}\Big(T^{-\frac{1}{2}}(\mathbf{\Psi})\theta, I\Big),$$

where $T^{-1}(\mathbf{\Psi}) := \Phi^T S^{-1}(\mathbf{\Psi})\Phi$ and $T^{-\frac{1}{2}}(\mathbf{\Psi})$ are symmetric matrices such that $T^{-1} = T^{-\frac{1}{2}}T^{-\frac{T}{2}}$. Denote the dimension of the estimated parameter n_θ. Since $\hat{\theta}^{\mathrm{N}}$ is Gaussian (at least asymptotically) a χ^2-test can be constructed to test between $\mathcal{H}_0$ and $\mathcal{H}_1$ using

$$\|\hat{\theta}^{\mathrm{N}}\|_2^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma'. \quad (16a)$$

For this test, called the *Wald test* [1], the asymptotic statistics becomes:

$$\|\hat{\theta}^{\mathrm{N}}\|_2^2 \overset{a}{\sim} \begin{cases} \chi^2_{n_\theta}, & \text{under } \mathcal{H}_0 \\ \chi'^2_{n_\theta}(\lambda), & \text{under } \mathcal{H}_1 \end{cases}, \quad (16b)$$

where $\chi'^2_{n_\theta}(\lambda)$ is the non-central χ^2-distribution with the non-centrality parameter

$$\lambda = \theta_1^T T^{-1}(\mathbf{\Psi})\theta_1 = \theta_1^T \Phi^T S^{-1}(\mathbf{\Psi})\Phi\theta_1, \quad (16c)$$

where θ_1 is the true parameter under $\mathcal{H}_1$.

B. Asymptotic GLR Test

It is possible to use the *generalized likelihood ratio* (GLR) test for faults (changes). If the PDF $p(\mathbb{Y}|\theta)$ is known, except for θ under $\mathcal{H}_1$ for hypotheses defined as in (2), then a detector can be constructed using the decision rule

$$L_G(\mathbb{Y}) = \frac{p(\mathbb{Y}|\hat{\theta}_1)}{p(\mathbb{Y}|\theta = 0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma,$$

where $\hat{\theta}_1$ is the MLE estimate of θ under $\mathcal{H}_1$. This is the GLR test.

A reason to use GLR is that it is known to be an asymptotically *uniformly most powerful* (UMP) test amongst all *invariant* tests, according to the theorem below. The GLR test is also known to be optimal for many special cases [15].

Theorem 2. *The* GLR *test is asymptotically* UMP, i.e.*, most powerful for all θ under $\mathcal{H}_1$, amongst all tests that are* invariant*. Furthermore, the asymptotic statistics are given by*

$$L'_G(\mathbb{Y}) := 2\log L_G(\mathbb{Y}) \overset{a}{\sim} \begin{cases} \chi^2_{n_\theta}, & \textit{under } \mathcal{H}_0 \\ \chi'^2_{n_\theta}(\lambda), & \textit{under } \mathcal{H}_1 \end{cases},$$

where

$$\lambda = \theta_1^T \mathcal{I}(\theta = 0)\theta_1.$$

The dimension of $\mathcal{I}(\theta = 0)$ is $n_\theta \times n_\theta$ and θ_1 is the true value of θ under the hypothesis $\mathcal{H}_1$.

Proof. See [1, Ch. 6]. □

It is shown in [1] that the Wald test has the same asymptotic properties as the generalized likelihood ratio test, and hence that it is asymptotically UMP amongst all invariant tests. For anything but theoretical argumentation, the assumption of infinitely many measurements is unrealistic. However, the asymptotic behavior constitutes a fundamental upper performance limit and as such it indicates how much better performance could be hoped for utilizing available information about non-Gaussian noise. Furthermore, in practice the GLR test usually performs quite well for moderately sized series of measurements [1]. Hence, the asymptotic behavior indicates what kind of performance to expect.

C. Detection Performance

Using one of the tests described above for a fixed threshold γ', the *probability of a false alarm*, P_{FA}, can be calculated as

$$P_{\mathrm{FA}} = \mathcal{Q}_{\chi^2_{n_\theta}}(\gamma'), \quad (17)$$

where $\mathcal{Q}_\star$ denotes the complementary cumulative density function of the distribution $\star$. Note that, P_{FA} depends only on the choice of threshold, γ', hence any change in noise distributions will only affect the *probability of detection*,

$$P_{\mathrm{D}} = \mathcal{Q}_{\chi'^2_{n_\theta}(\lambda)}(\gamma'), \quad (18)$$

where λ is defined by (16c). The $\mathcal{Q}_{\chi'^2_{n_\theta}(\lambda)}(\gamma')$ is monotonously increasing in λ (moving the mean to the left lessens the risk that a detection is missed) thus any increase in λ will improve P_{D}. It follows immediately from (16c) that if the magnitude of θ_1 increases it is easier to detect the change. Further, if S is increased P_{D} increases, too. From the expressions for S in (11) it is clear that any increase in RA increases λ, and since $\Psi > 1$ for non-Gaussian noise it follows that any non-Gaussian noise improves P_{D} compared to the same system with Gaussian noise.

Example. Consider measurements from

$$y_t = \theta + e_t, \quad t = 1, \ldots, L, \quad (19)$$

where e_t is any noise with $\mathsf{var}(e) = 1$ and Ψ_e quantifying any non-Gaussian noise properties. It then follows that

$$\hat{\theta}^{\mathrm{N}} \overset{a}{\sim} \mathcal{N}\Big(\sqrt{\Psi L}\theta, I\Big),$$

that subsequently leads to $\lambda = \Psi_e L\theta_1^2 = \Psi_e L$ assuming $\theta_1 = 1$. The improved detection performance is illustrated in

Fig. 1 by the *receiver operating characteristics* (ROC). From the figure it is clear that there is potentially much to be gained from utilizing information about non-Gaussian noise in this simple model, especially for small P_{FA} where the relative increase in P_{D} is quite substantial. $\diamond$

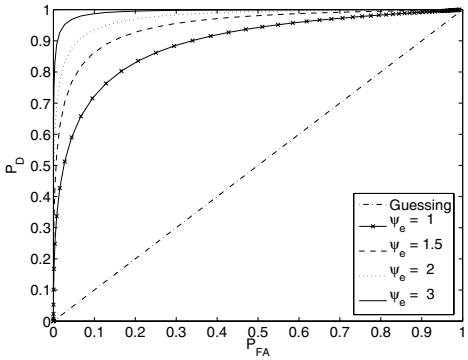

Fig. 1. ROC diagram for (19) with $n_\theta = 1$, $L = 5$ measurements, and $\theta_1 = 1$ for noise with different Ψ_e values. Guessing denotes what happens if all available information is discarded and a change is signaled randomly with probability P_{FA}, it is always possible to construct a test at least this good.

IV. Application: Fault Detection with Outliers

Consider again measurements from (19) where this time e_t is Gaussian measurement noise affected by outliers. The outliers result in heavier tails in the PDF than in a Gaussian PDF. The noise can be modeled as a *Gaussian mixture*,

$$e \sim (1-\omega)\mathcal{N}(0,R) + \omega\mathcal{N}(0,kR), \tag{20}$$

where $0 \leq \omega \leq 1$ is the probability of outliers, k indicates how many times larger the variance of the outliers is, and R is the variance of measurements unaffected by outliers. Fig. 2 shows the PDF of noise with $\mathsf{var}(e) = 1$ ($R = 0.277$), $\omega = 0.1$, and $k = 10$. This distribution has $\Psi_e = 1.5$.

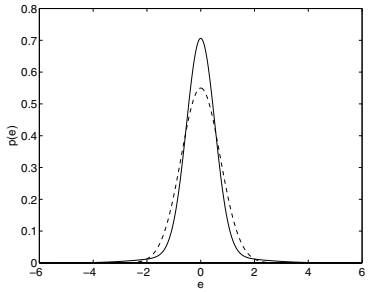

Fig. 2. Solid line, PDF for measurement noise with outliers, parameterized as in (20), for $\omega = 0.1$, $k = 10$, $\mathsf{var}(e) = 1$, and $\Psi_e = 1.5$. Dashed line shows the Gaussian approximation.

Trying to detect a change in θ the performance limits derived above apply. First assuming Gaussian noise with correct variance (hence a linear approximation) yields for $P_{\text{FA}} = 1\%$ a probability of detection $P_{\text{D}} = 37\%$, assuming $\theta_1 = 1$ and $L = 5$.

The probability of detection can be increased by utilizing information about outliers in the measurements since

$$P_{\text{FA}} = \mathcal{Q}_{\chi^2_1}(\gamma')$$

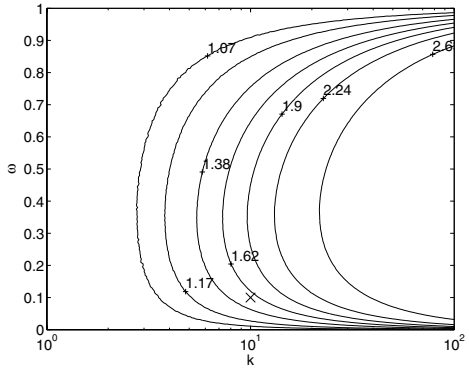

Fig. 3. Normalized probability of detection P_{D} in noise with outliers (20), $\mathsf{var}(e) = 1$, and $P_{\text{FA}} = 0.01$. The level curves are normalized so that 1 denotes $P_{\text{D}} = 0.37$, *i.e.*, what is achieved for Gaussian noise, 1.62 denotes $P_{\text{D}} = 1.62 \cdot 0.37 = 0.60$ *etc.*

is independent of Ψ_e but

$$P_{\text{D}} = \mathcal{Q}_{\chi'^2_1(\lambda)}(\gamma') = \mathcal{Q}_{\chi'^2_1(\Psi_e L)}(\gamma')$$

is not. The improvement that comes from the Ψ_e dependency is shown in Fig. 3. As can be seen, most detectability is gained with moderately many outliers, $\omega \approx 30\%$, with large variance, $k \gg 1$, since this results in a large relative increase in P_{D}. The situation with 10% outliers ($\omega = 0.1$) of 10 times the variance ($k = 10$) mentioned above, denoted with $\times$ in Fig. 3, results in a relative $P_{\text{D}} = 157\%$. The chance of detection is improved with more than 55%.

For the same system, (19) with the same measurement noise (20) ($\omega = 0.1$, $k = 10$, and $\mathsf{var}(e) = 1$ denoted with $\times$ in Fig. 3), *Monte Carlo* (MC) simulations have been carried out to show how close the ROC diagram comes to the optimal curve. With a GLR test based on 5 measurements from (19), numerically computed MLE of $\hat{\theta}$, and 10 000 MC simulations, Fig. 4 is achieved. The result is promising since the simulations seem to come close to the performance bound. Note that Fig. 1 and Fig. 4 show the same relation derived analytically and from simulations, respectively.

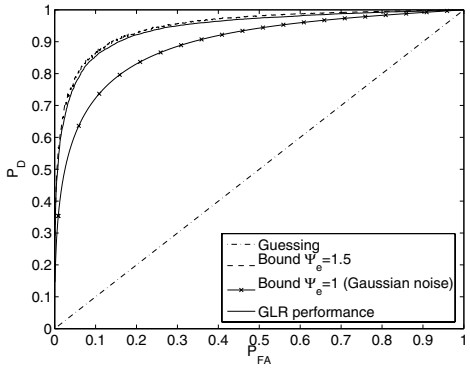

Fig. 4. ROC diagram for (19) with measurements (20) ($\omega = 0.1$, $k = 10$, $\mathsf{var}(e) = 1$, and $\Psi_e = 1.5$) for 10 000 MC simulations. Optimal performance, also found in Fig. 1, is included as reference for $\Psi_e = 1$ and $\Psi_e = 1.5$.

V. Conclusion

Optimal detection performance in Gaussian and non-Gaussian noise has been studied for linear regression residuals; explicit calculations have been performed for linear regression, and dynamic linear systems with estimated initial state as well as a parity space formulation. Detection performance in terms of probability of detection, P_{D}, for a given probability of false alarm, P_{FA}, is expressed in terms of *intrinsic accuracy* (IA) and *relative accuracy* (RA). Using these results, it is possible to decide if more advanced, computationally expensive, methods for detection should be evaluated. Monte Carlo simulations show that for a moderately sized window of measurements it is possible to come close to optimal performance in a situation with outliers in the measurements.

Appendix I
Definition of Derivatives

The derivative of $f : \mathbb{R}^n \mapsto \mathbb{R}^m$, often denoted the *gradient* or the *Jacobian*, used is

$$\nabla_x f = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}.$$

With this definition the second derivative of a scalar function $f : \mathbb{R}^n \mapsto \mathbb{R}$, also called *Hessian* when $x = y$, becomes

$$\Delta_x^y f = \nabla_x \nabla_y f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial y_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial y_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial y_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial y_m} \end{pmatrix}.$$

Appendix II
Properties of Intrinsic Accuracy

Lemma 1. *For a vector* $\mathbb{E}$ *of* independently and identically distributed *(*iid*) random variables* $\mathbb{E} = (e_1^T, \ldots, e_n^T)^T$ *each with* $\mathsf{cov}(e_i) = R$ *and* $\mathcal{I}_{e_i} = \mathcal{I}_e$*,*

$$\mathsf{cov}(\mathbb{E}) = I \otimes R \quad \textit{and} \quad \mathcal{I}_{\mathbb{E}} = I \otimes \mathcal{I}_e,$$

where $\otimes$ *denotes the* Kronecker product. *If* $R\mathcal{I}_e = \Psi_e \cdot I$*, with* $\Psi_e \geq 1$ *a scalar, then*

$$\mathsf{cov}(\mathbb{E}) = \Psi_e \mathcal{I}_{\mathbb{E}}^{-1} = \Psi_e I \otimes \mathcal{I}_e^{-1}.$$

Proof. For $\mathbb{E} = (e_1^T, e_2^T, \ldots, e_n^T)^T$, e_i iid,

$$\begin{aligned} \mathcal{I}_{\mathbb{E}} &= -\mathsf{E}\,\Delta_{\mathbb{E}}^{\mathbb{E}} \log p(\mathbb{E}) \\ &= -\mathsf{E}\,\Delta_{\mathbb{E}}^{\mathbb{E}} \log \prod_{i=1}^{n} p(e_i) = \sum_{i=1}^{n} -\mathsf{E}\,\Delta_{\mathbb{E}}^{\mathbb{E}} \log p(e_i). \end{aligned}$$

For this the derivatives becomes, for $k = l = i$

$$-\mathsf{E}\,\nabla_{e_k}\nabla_{e_l} \log p(e_i) = -\mathsf{E}\,\Delta_{e_i}^{e_i} \log p(e_i) = \mathcal{I}_e,$$

and otherwise $-\mathsf{E}\,\nabla_{e_k}\nabla_{e_l} \log p(e_i) = 0$. Combining these to get back to matrix notation yields

$$\mathcal{I}_{\mathbb{E}} = \operatorname{diag}(\mathcal{I}_{e_i}) = I \otimes \mathcal{I}_e.$$

This concludes the proof of Lemma 1. □

Using [18], the following theorem can be shown.

Theorem 3. *For* $\mathbb{Z} = B\mathbb{E}$*, where* $\mathbb{E} = (e_1^T, \ldots, e_n^T)^T$ *is a stochastic variable with* iid *components such that* $\mathsf{cov}(e_i) = R$ *and* $\mathcal{I}_{e_i} = \mathcal{I}_e$ *then*

$$\begin{aligned} \mathsf{cov}(\mathbb{Z}) &= B\big(I \otimes R\big)B^T, \\ \mathcal{I}_{\mathbb{Z}}^{-1} &= B\big(I \otimes \mathcal{I}_e^{-1}\big)B^T. \end{aligned}$$

Furthermore, if Ψ_e *is relative accuracy for* e_i *then*

$$\mathsf{cov}(\mathbb{Z}) = \Psi_e \mathcal{I}_{\mathbb{Z}}^{-1} = \Psi_e B\big(I \otimes \mathcal{I}_e^{-1}\big)B^T.$$

Proof. Combine the result found as Theorem 4.3 in [18] with Lemma 1. For the last property use Definition 2. □

Acknowledgment

This work is supported by VINNOVA's Center of Excellence ISIS (Information Systems for Industrial Control and Supervision) at Linköpings universitet, Linköping, Sweden.

References

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice-Hall, Inc, 1998, vol. 2.

[2] E. L. Lehmann, *Testing Statistical Hypotheses*, 2nd ed., ser. Probability and mathematical Statistics. John Wiley & Sons, Ltd, 1986.

[3] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gausian Bayesian state estimation," *IEE Proc.-F*, vol. 140, no. 2, pp. 107–113, Apr. 1993.

[4] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, ser. Statistics for Engineering and Information Science. New York: Springer-Verlag, 2001.

[5] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, Inc, 2004.

[6] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc, 1993, vol. 1.

[7] E. L. Lehmann, *Theory of Point Estimation*, ser. Probability and mathematical Statistics. John Wiley & Sons, Ltd, 1983.

[8] S. M. Kay and D. Sengupta, "Optimal detection in colored non-Gaussian noise with unknown parameter," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, vol. 12, Dallas, TX, USA, Apr. 1987, pp. 1087–1089.

[9] D. Sengupta and S. M. Kay, "Parameter estimation and GLRT detection in colored non-Gaussian autoregressive processes," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 10, no. 38, pp. 1661–1676, Oct. 1990.

[10] G. Hendeby and F. Gustafsson, "Fundamental filtering limitations in linear non-Gaussian systems," in *Proc. 16th Triennial IFAC World Congress*, Prague, Czech Republic, July 2005.

[11] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*. New York: Chapman and Hall, 1974.

[12] D. Sengupta and S. M. Kay, "Efficient estimation for non-Gaussian autoregressive processes," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 6, pp. 785–794, June 1989.

[13] F. Gustafsson, "Stochastic fault diagnosability in parity spaces," in *Proc. 15th Triennial IFAC World Congress on Automatic Control*, Barcelona, Spain, July 2005.

[14] D. Törnqvist, F. Gustafsson, and I. Klein, "GLR tests for fault detection over sliding data windows," in *Proc. 16th Triennial IFAC World Congress*, Prague, Czech Republic, July 2005.

[15] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Inc, 1993.

[16] J. J. Gertler, *Fault Detection and Diagnosis in Engineering Systems*. Marcel Dekker, Inc, 1998.

[17] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. John Hopkins University Press, 1996.

[18] N. Bergman, "Recursive Bayesian estimation: Navigation and tracking applications," Dissertations No 579, Linköping Studies in Science and Technology, SE-581 83 Linköping, Sweden, May 1999.