DATA VALIDATION IN THE PRESENCE OF IMPRECISELY KNOWN CORRELATIONS

U. D. Hanebeck[1], **J. Horn**[2]

[1]Inst. of Computer Design and Fault Tolerance, Universität Karlsruhe, 76128 Karlsruhe, Germany, Uwe.Hanebeck@ieee.org
[2]Siemens AG, Corporate Technology, Information and Comm., 81730 München, Germany, Joachim.P.Horn@siemens.com

Abstract

This paper derives fundamental results for data validation in the presence of imprecisely known correlations. Given a constraint on the maximum absolute correlation of a given estimate and measurement data, a tight upper bound for the joint covariance matrix is derived, which finally yields a modified Mahalanobis distance. The special cases of one–dimensional and two–dimensional random variables are discussed.

Keywords: Data Validation; Stochastic Uncertainties; Mahalanobis Distance; Imprecisely Known Correlations; Covariance Bounds

1 Introduction

Data validation is a crucial task in most filtering applications: Given an estimate of the unknown state, a noisy measurement, and a quantification of the uncertainty of the estimate and the measurement data, data validation has to decide if the measurement data is compatible with the given estimate, and thus can be used for filtering. Stochastic data validation for linear systems corrupted by Gaussian noise is typically based on the Mahalanobis distance [1]. For calculating the Mahalanobis distance, the cross–covariance of the given estimate and the measurement data has to be known. However, in many applications, the cross–covariance is not known precisely. Only constraints on the maximum correlation coefficient are given.

This paper derives a modified Mahalanobis distance that allows for data validation when only constraints on the maximum correlation coefficient are known. It is based on a tight bound for the joint covariance of two random vectors with unknown but constrained cross–correlation derived in [3]. Applications of this data validation scheme include data association problems in simultaneous vehicle localization and map building [6].

2 Problem Formulation

We are given two random vectors $\underline{x} \in \mathbb{R}^N$, $\underline{y} \in \mathbb{R}^M$ with expected values

$$E\{\underline{x}\} = \hat{\underline{x}} \ , \ E\{\underline{y}\} = \hat{\underline{y}}$$

and individual covariances

$$\mathrm{Cov}\{\underline{x}\} = \mathbf{C}_{xx} \ , \ \mathrm{Cov}\{\underline{y}\} = \mathbf{C}_{yy} \ ,$$

where $\underline{x}$ and $\underline{y}$ are assumed to be correlated. Their cross covariances $\mathrm{Cov}\{\underline{x}, \underline{y}\} = \mathbf{C}_{xy}$ and $\mathrm{Cov}\{\underline{y}, \underline{x}\} = \mathbf{C}_{yx}$, however, are not explicitly known. Correlations between $\underline{x}$ and $\underline{y}$ are modelled by a total correlation coefficient r_{xy}, which is constrained according to

$$|r_{xy}| \leq r_{max} \ . \tag{1}$$

Of course, this also includes the case of completely unknown correlation between $\underline{x}$ and $\underline{y}$ for $r_{max} = 1$.

Hence, a constraint for the cross covariances is given by

$$\mathbf{C}_{yx} \, \mathbf{C}_{xx}^{-1} \, \mathbf{C}_{xy} \leq r_{max}^2 \mathbf{C}_{yy} \ , \tag{2}$$

where, in general, for two positive definite matrices $\mathbf{A}$ and $\mathbf{B}$, an expression of the form $\mathbf{A} > \mathbf{B}$ ($\mathbf{A} \geq \mathbf{B}$) is interpreted as $\mathbf{A} - \mathbf{B}$ positive definite (positive semi–definite). By defining the matrix

$$\mathbf{C} = r_{max}^2 \mathbf{C}_{yy} - \mathbf{C}_{yx} \, \mathbf{C}_{xx}^{-1} \, \mathbf{C}_{xy} \ ,$$

(2) is equivalent to

$$\det\left(\mathbf{C}(1:i, 1:i)\right) \geq 0$$

for $i = 1, \ldots, M$ according to the Sylvester criterion.

We now assume that an estimate of $\underline{x}$ is available and new information in the form of measurement data $\underline{y}$ is obtained. We also assume that the (unknown) true vectors $\tilde{\underline{x}}$ and $\tilde{\underline{y}}$ are related by a measurement equation of the form

$$\mathbf{H}_x \, \tilde{\underline{x}} = \mathbf{H}_y \, \tilde{\underline{y}} \ , \tag{3}$$

where $\mathbf{H}_x$ and $\mathbf{H}_y$ are known matrices of appropriate dimensions.

Now, the following question arises: Is $\underline{y}$ a valid measurement, i.e., can it be explained by the measurement equation and be used for estimation purposes or is it an outlier? Of course, this question can only be answered with a certain probability of, say 99 %, of being true. The task of data validation is complicated by the fact, that the correlation between $\underline{x}$ and $\underline{y}$ is not known precisely.

The problem is now solved in two steps. In the first step, covariance bounds for two random variables with imprecisely known correlation are derived. In the second step, these bounds are used to derive a modified Mahalanobis distance for data validation in the presence of imprecisely known correlations.

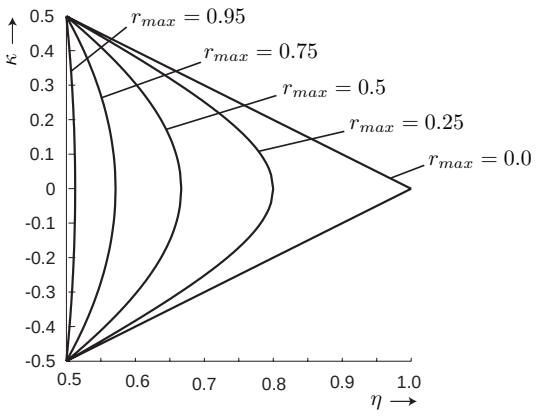

Figure 1: The admissible values for $\eta(\kappa)$ and κ resulting from Lemma 3.1.

3 Covariance Bounds

In this section, we are concerned with deriving covariance bounds for the joint covariance of two correlated random variables with imprecisely known correlation. Hence, our goal is now to find a family of bounding covariances $\mathbf{B}$ with

$$\mathbf{B} \geq \mathbf{C}(r_{xy}) \tag{4}$$

for all possible joint covariances $\mathbf{C}(r_{xy})$ defined by

$$\mathbf{C}(r_{xy}) = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix}$$

with r_{xy} according to (1) and $\mathbf{C}_{xy}$, $\mathbf{C}_{yx}$ such that (2) holds.

For deriving the desired covariance bounds we use the fact that the union of the 1-sigma–bounds of all possible joint covariances forms a convex set aligned with the coordinate axes. Hence, the cross covariances of the bounding covariance matrix have to be zero matrices. For the simplest case of two scalar random variables x and y this is visualized in Figure 2.

In addition, for achieving an upper bound, the covariance matrices $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ have to be individually scaled. Combining both conditions yields

$$\mathbf{B} = \begin{bmatrix} k_x \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & k_y \mathbf{C}_{yy} \end{bmatrix} \ . \tag{5}$$

k_x, k_y have to be selected in such a way that (4) holds.

THEOREM 3.1 *The scale factors k_x, k_y in (5) are given by*

$$k_x = \frac{1}{\eta - \kappa}\ ,\ k_y = \frac{1}{\eta + \kappa} \tag{6}$$

with

$$\kappa^2 \leq \frac{1 - 2\eta}{1 - r_{max}^2} + \eta^2 \tag{7}$$

and

$$0.5 \leq \eta \leq \frac{1}{1 + r_{max}} \ . \tag{8}$$

PROOF. For proving (4), the difference matrix

$$\begin{aligned} \mathbf{D} &= \mathbf{B}(\eta, \kappa) - \mathbf{C}(r_{xy}) \\ &= \begin{bmatrix} \frac{1}{\eta - \kappa}\mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\eta + \kappa}\mathbf{C}_{yy} \end{bmatrix} - \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix} \\ &= \begin{bmatrix} \left(\frac{1}{\eta - \kappa} - 1\right)\mathbf{C}_{xx} & -\mathbf{C}_{xy} \\ -\mathbf{C}_{yx} & \left(\frac{1}{\eta + \kappa} - 1\right)\mathbf{C}_{yy} \end{bmatrix} \end{aligned}$$

is considered. According to Sylvester's criterion, the matrix $\mathbf{D}$ is positive semi–definite, if the determinants of all submatrices $\mathbf{D}(1 : N + i, 1 : N + i)$ for $i = 1, \ldots, M$ are larger than or equal to zero. $\left(\frac{1}{\eta - \kappa} - 1\right)\mathbf{C}_{xx}$ is positive definite and does not need to be tested. The determinants are given by

$$\begin{aligned} &\det(\mathbf{D}(1 : N + i, 1 : N + i)) \\ &= \det\left(\left(\frac{1}{\eta - \kappa} - 1\right)\mathbf{C}_{xx}\right) \det\left(\left(\frac{1}{\eta + \kappa} - 1\right)\mathbf{C}_{yy}(1 : i, 1 : i)\right. \\ &\quad \left. - \left(\frac{1}{\eta - \kappa} - 1\right)^{-1} \mathbf{C}_{yx}(1 : i, 1 : N)\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}(1 : N, 1 : i)\right) \ . \end{aligned}$$

With (2) we obtain

$$\det\left(\left[\left(\frac{1}{\eta - \kappa} - 1\right)\left(\frac{1}{\eta + \kappa} - 1\right) - r_{max}^2\right]\mathbf{C}_{yy}(1 : i, 1 : i)\right) \stackrel{!}{\geq} 0$$

for $i = 1, \ldots, M$, which is equivalent to

$$\left(\frac{1}{\eta - \kappa} - 1\right)\left(\frac{1}{\eta + \kappa} - 1\right) - r_{max}^2 \geq 0$$

and yields (7). The constraint on η in (8) then follows by claiming a non–negative right–hand–side in (7). □

The parameter set for η and κ from the Theorem is redundant in the sense that it specifies scaled variants of a bounding covariance with the same form and orientation. Hence, it is sufficient to restrict attention to the smallest of these scaled variants. The appropriate parameter values are specified in the following Lemma.

LEMMA 3.1 *A family of bounding covariances $\mathbf{B}(\kappa)$ depending on a parameter κ is given by (5) with k_x, k_y in (6). The parameter κ may vary according to*

$$|\kappa| \leq 0.5 \ . \tag{9}$$

η is a function of κ given by

$$\eta(\kappa) = \frac{1 - \sqrt{r_{max}^2 + \kappa^2\left(1 - r_{max}^2\right)^2}}{1 - r_{max}^2} \ . \tag{10}$$

The admissible values for η and κ resulting from Lemma 3.1 are visualized for different values of r_{max} in Figure 1.

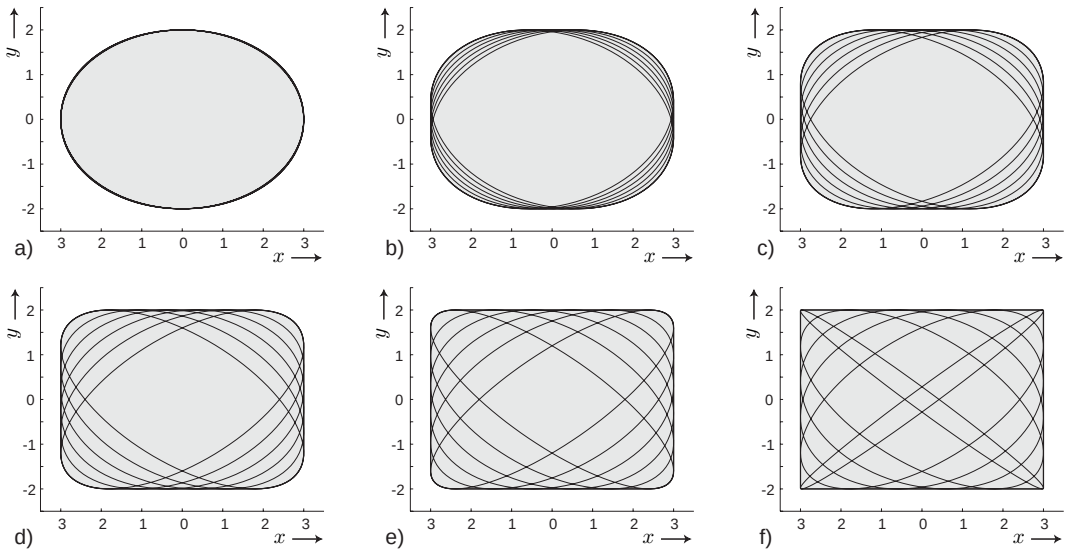

Figure 2: Union of the ellipses corresponding to the joint covariances of two scalar random variables x and y with imprecisely known correlation in Example 4.1: a) $|r_{xy}| < 0.01$, b) $|r_{xy}| < 0.2$, c) $|r_{xy}| < 0.4$, d) $|r_{xy}| < 0.6$, e) $|r_{xy}| < 0.8$, f) $|r_{xy}| < 0.99$. The unconstrained case corresponds to $|r_{xy}| \leq 1$.

4 Data Validation

In the case of precisely known correlation, we define the Mahalanobis distance

$$d = (\mathbf{H}_x\,\underline{x} - \mathbf{H}_y\,\underline{y})^T \left(\mathrm{Cov}\{\mathbf{H}_x\,\underline{x} - \mathbf{H}_y\,\underline{y}\}\right)^{-1} (\mathbf{H}_x\,\underline{x} - \mathbf{H}_y\,\underline{y}) \ ,$$

with

$$\begin{aligned}\mathrm{Cov}\{\mathbf{H}_x\,\underline{x} - \mathbf{H}_y\,\underline{y}\} = \ & \mathbf{H}_x\,\mathbf{C}_{xx}\,\mathbf{H}_x^T + \mathbf{H}_y\,\mathbf{C}_{yy}\,\mathbf{H}_y^T \\ & - \mathbf{H}_x\,\mathbf{C}_{xy}\,\mathbf{H}_y^T - \mathbf{H}_y\,\mathbf{C}_{yx}\,\mathbf{H}_x^T \ .\end{aligned}$$

However, in the case of uncertain correlation between $\underline{x}$ and $\underline{y}$, the matrices $\mathbf{C}_{xy}$ and $\mathbf{C}_{yx}$ are not known precisely. Hence, the Mahalanobis distance depends upon the uncertain correlation between the given estimate $\underline{x}$ and the measurement $\underline{y}$.

An efficient data validation procedure is now derived by replacing the covariance $\mathrm{Cov}\{\mathbf{H}_x\,\underline{x} - \mathbf{H}_y\,\underline{y}\}$ by the upper covariance bounds derived before. Then we have

$$\begin{aligned}d \geq d(\kappa) = & (\mathbf{H}_x\,\underline{x} - \mathbf{H}_y\,\underline{y})^T \left(\frac{\mathbf{H}_x\,\mathbf{C}_{xx}\,\mathbf{H}_x^T}{\eta(\kappa) - \kappa} \right. \\ & \left. + \frac{\mathbf{H}_y\,\mathbf{C}_{yy}\,\mathbf{H}_y^T}{\eta(\kappa) + \kappa} \right)^{-1} (\mathbf{H}_x\,\underline{x} - \mathbf{H}_y\,\underline{y}) \ .\end{aligned}$$

Hence, a lower bound $d(\kappa)$ for the true Mahalanobis distance d is obtained. However, it can be shown that this bound actually attains the true bound for some value of $\kappa \in (-0.5, 0.5)$. The proof is outside the scope of this paper. This fact allows the application of the modified Mahalanobis distance to perform an *exact* data validation according to the above assumptions.

A given measurement $\underline{y}$ is accepted, if

$$\max_{\kappa \in (-0.5, 0.5)} d(\kappa) < k \ ,$$

where k is a given gating threshold.

It is important to note that $d(\kappa)$ is a concave function. Hence, the (unique) maximum can quickly be found by, for example, bisection routines. Analytic solutions for the maximum value are also available for certain special cases.

Data validation in the case of completely unknown correlations is similar to testing the overlap of two ellipsoids, which has been treated in [5, 8]. Efficient analytic solutions based on the ideas in [7] are given in [2].

4.1 Special Case: One–dimensional Random Variables

We now consider scalar random variables x and y with variances C_{xx} and C_{yy}, respectively.

EXAMPLE 4.1 For two scalar random variables x and y with individual variances $C_{xx} = 9$ and $C_{yy} = 4$, some members of the family of possible joint covariance matrices for different constraints on the maximum absolute correlation coefficient are visualized in Figure 2 by plotting the respective 1–sigma–bounds.

The covariance bound $\mathbf{B}(\kappa)$ for the true joint covariance $\mathbf{C}(r_{xy})$ is given by

$$\mathbf{B}(\kappa) = \begin{bmatrix} \frac{1}{\eta(\kappa) - \kappa} C_{xx} & 0 \\ 0 & \frac{1}{\eta(\kappa) + \kappa} C_{yy} \end{bmatrix} \ .$$

EXAMPLE 4.2 The covariance bounds for the two scalar random variables in Example 4.1 are shown in Figure 3.

With

$$H_x\,x = H_y\,y$$

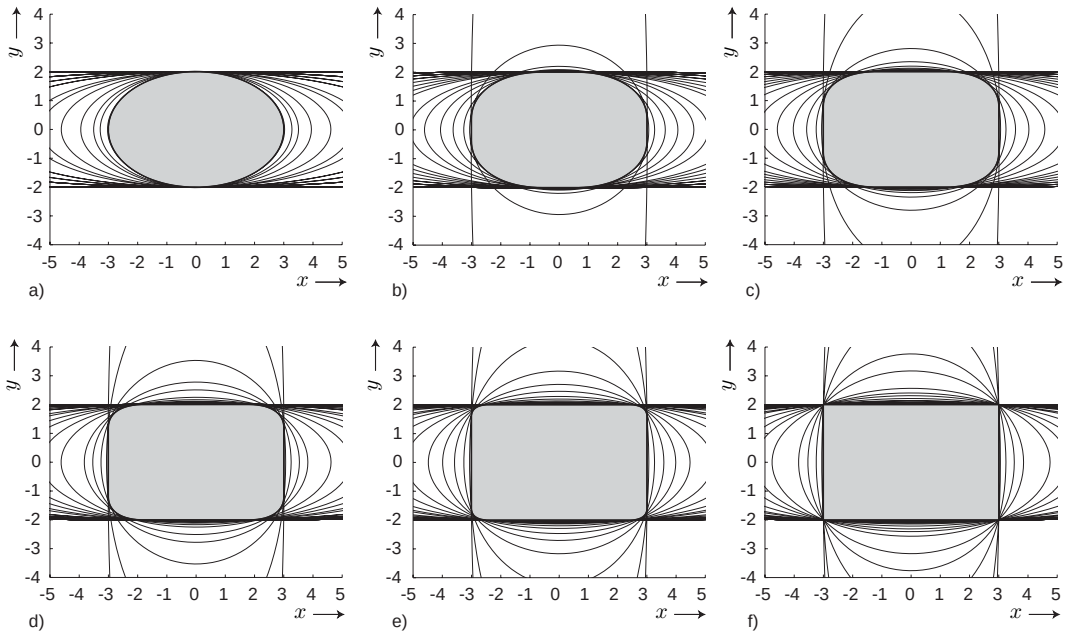

Figure 3: Covariance bounds for the joint covariances of two scalar random variables x and y with imprecisely known correlation in Figure 2: a) $|r_{xy}| < 0.01$, b) $|r_{xy}| < 0.2$, c) $|r_{xy}| < 0.4$, d) $|r_{xy}| < 0.6$, e) $|r_{xy}| < 0.8$, f) $|r_{xy}| < 0.99$. The unconstrained case corresponds to $|r_{xy}| \leq 1$.

the modified Mahalanobis distance is given by

$$d(\kappa) = \frac{\left(\eta^2(\kappa) - \kappa^2\right)\,(H_x\,x - H_y\,y)^2}{(\eta(\kappa) + \kappa)\;H_x^2\,C_{xx} + (\eta(\kappa) - \kappa)\;H_y^2\,C_{yy}} \quad .$$

REMARK 4.1 In the case of one–dimensional random variables, the optimal κ can be calculated without knowledge of the actual measured values x and y. It just depends upon the actual values of the variances C_{xx} and C_{yy}.

4.2 Special Case: Two–dimensional Random Variables

We consider two two–dimensional random vectors $\underline{x} \in \mathbb{R}^2$, $\underline{y} \in \mathbb{R}^2$. The individual covariance matrices $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are assumed to be known. The correlation between $\underline{x}$ and $\underline{y}$ is not known precisely and characterized by a total correlation coefficient constrained according to $|r_{xy}| \leq r_{max}$.

A simple method for generating valid joint covariance matrices

$$\mathrm{Cov}\left\{\begin{bmatrix}\underline{x}\\ \underline{y}\end{bmatrix}\right\} = \begin{bmatrix}\mathbf{C}_{xx} & \mathbf{C}_{xy}\\ \mathbf{C}_{yx} & \mathbf{C}_{yy}\end{bmatrix}$$

works as follows. $\mathbf{C}_{xy}$ with $\mathbf{C}_{xy} = \mathbf{C}_{yx}^T$ contains four unknown elements. These elements are generated based on individual correlation coefficients defined by

$$r_{11} = \frac{\mathbf{C}_{xy,11}}{\sqrt{\mathbf{C}_{xx,11}\,\mathbf{C}_{yy,11}}} \quad , \qquad r_{12} = \frac{\mathbf{C}_{xy,12}}{\sqrt{\mathbf{C}_{xx,11}\,\mathbf{C}_{yy,22}}} \quad ,$$

$$r_{21} = \frac{\mathbf{C}_{xy,21}}{\sqrt{\mathbf{C}_{xx,22}\,\mathbf{C}_{yy,11}}} \quad , \qquad r_{22} = \frac{\mathbf{C}_{xy,22}}{\sqrt{\mathbf{C}_{xx,22}\,\mathbf{E}_{yy,22}}} \quad ,$$

which may vary according to $r_{11}, r_{12}, r_{21}, r_{22} \in [-r_{max}, r_{max}]$. Random values for these correlation coefficients are generated which are uniformly distributed in the interval $[-r_{max}, r_{max}]$. Based upon these values, tentative entries of the cross covariance matrices $\mathbf{C}_{xy}$ and $\mathbf{C}_{yx}$ are calculated. These tentative entries are validated by means of (2). The method is summarized in Figure 4.

EXAMPLE 4.3 We consider two two–dimensional zero–mean random vectors $\underline{x}$ and $\underline{y}$ with individual covariances

$$\mathbf{C}_{xx} = \begin{bmatrix}3 & -1\\ -1 & 1\end{bmatrix} \quad , \qquad \mathbf{C}_{yy} = \begin{bmatrix}5 & 4\\ 4 & 5\end{bmatrix} \quad ,$$

which are correlated with $|r_{xy}| \leq 0.8$.

In addition, the two random vectors are related according to (3) with

$$\mathbf{H}_x = \begin{bmatrix}4 & 0\\ 0 & -1\end{bmatrix}$$

and

$$\mathbf{H}_y = \begin{bmatrix}1 & 0\\ 0 & 3\end{bmatrix} \quad .$$

Some members of the family of possible joint covariance matrices for the random vector $\underline{z}$ with

$$\underline{z} = \mathbf{H}_x\,\underline{x} - \mathbf{H}_y\,\underline{y}$$

are visualized in Figure 5 a) by plotting the respective 1–sigma–bounds. Some members of the family of outer covariance bounds derived in this paper are shown in Figure 5 b). The

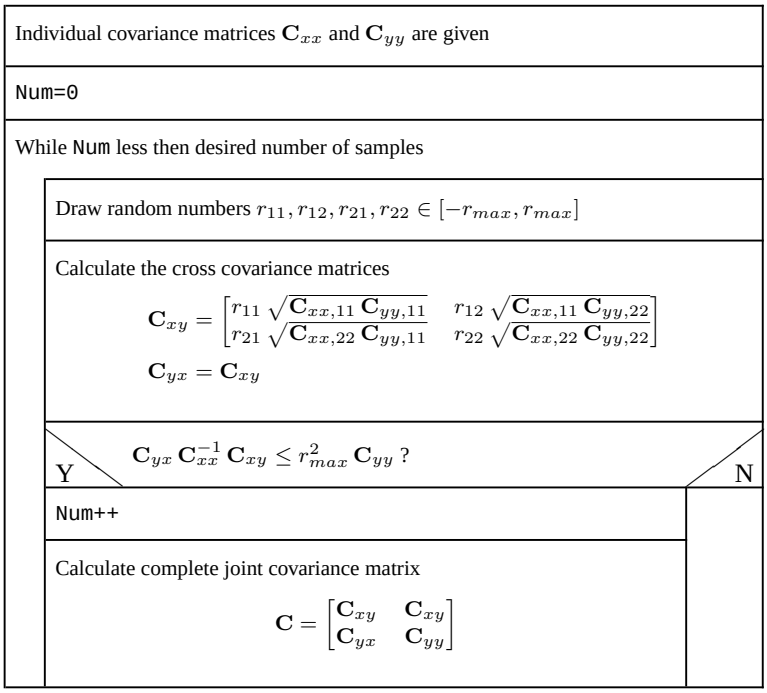

Figure 4: Structure chart of the naive generation of valid joint covariance matrices for two two–dimensional random vectors with symmetrically constrained correlation.

intersection of the outer covariance bounds is equivalent to the union of all possible covariances. This set is shown in Figure 6.

Data validation is now applied to samples generated for $\underline{x}$ and $\underline{y}$ with a gating threshold $k = 1$. The result is shown in Figure 7. Valid samples are shown in Figure 7 a). The invalid samples that are rejected by data validation are shown in Figure 7 b).

5 Conclusions

A new method for data validation in the presence of imprecisely known correlations has been proposed, which contains the case of completely unknown correlations [4] as a special case. It is exact, simple to implement, and computationally efficient. The main computational effort is spent by calculating the maximum of a scalar concave function.

References

[1] Y. Bar–Shalom and T. E. Fortmann, *Tracking and Data Association*, Academic Press, 1988.

[2] U. D. Hanebeck, *Lokalisierung eines mobilen Roboters mittels effizienter Auswertung von Sensordaten und mengenbasierter Zustandsschätzung*, PhD thesis, Lehrstuhl für Steuerungs– und Regelungstechnik, Technische Universität München, Fortschrittsberichte VDI, Reihe 8: Meß–, Steuerungs– und Regelungstechnik, Nr. 643, VDI Verlag, Düsseldorf, 1997.

[3] U. D. Hanebeck, K. Briechle, and J. Horn, "A Tight Bound for the Joint Covariance of Two Random Vectors with Unknown but Constrained Cross–Correlation", *Proceedings of the IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2001)*, 2001, pp. 85–90.

[4] U. D. Hanebeck and K. Briechle, "New Results for Stochastic Prediction and Filtering with Unknown Correlations", *Proceedings of the IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2001)*, 2001, pp. 147–152.

[5] T. H. Kerr, "Real–Time Failure Detection: A Static Non–Linear Optimization Problem That Yields a Two Ellipsoid Overlap Test", *Journal on Optimization Theory Applications*, Vol. 2, 1977, pp. 509–536.

[6] J. Neira and J. D. Tardós, "Data Association in Stochastic Mapping Using the Joint Compatibility Test", *IEEE Transactions on Robotics and Automation*, Vol. 17, No. 6, 2001, pp. 890–897.

[7] J. W. Perram and M. S. Wertheim, "Statistical Mechanics of Hard Ellipsoids. I. Overlap Algorithm and the Contact Function", *Journal of Computational Physics*, Vol. 58, 1985, pp. 409–416.

[8] A. Zolghadri, B. Bergeon, and M. Monsion, "A Two–Ellipsoid Overlap Test for On–line Failure Detection", *Automatica*, Vol. 29, No. 6, 1993, pp. 1517–1522.

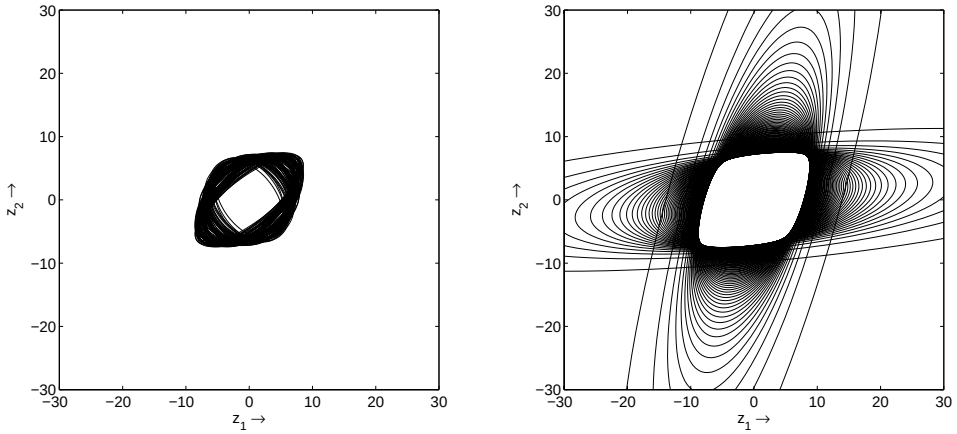

Figure 5: Visualizations for Example 4.3 a) (Left) Some members of the family of possible joint covariance matrices of $\mathbf{z}$ for the constraint $|r_{xy}| \leq 0.8$. b) (Right) Some members of the corresponding family of covariance bounds.

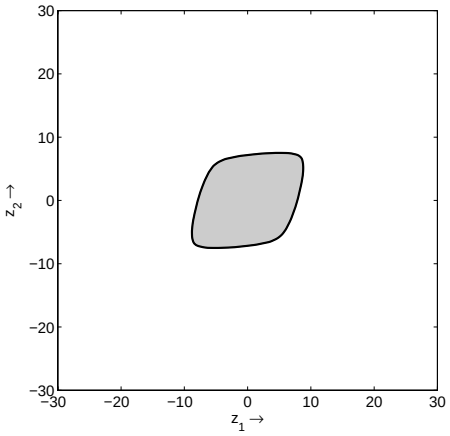

Figure 6: The set corresponding to the union of all possible covariances or equivalently to the intersection of all covariance bounds in Example 4.3.

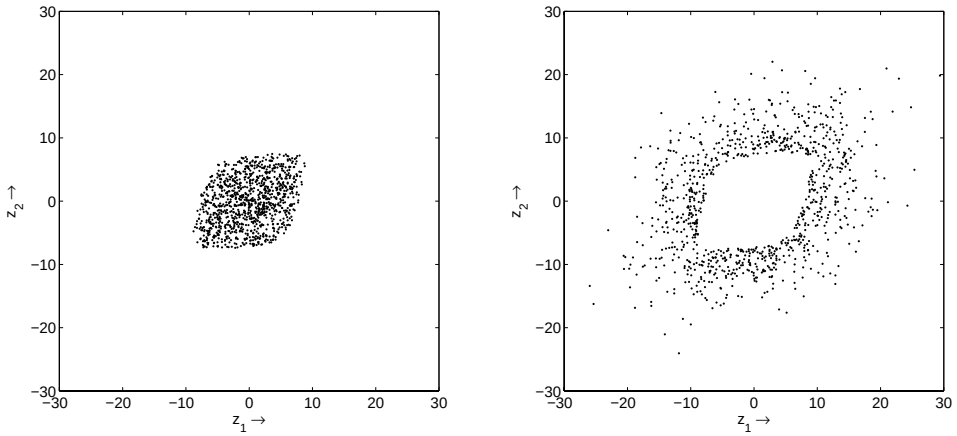

Figure 7: Visualizations for Example 4.3 a) (Left) Valid samples. b) (Right) Invalid samples rejected by data validation.