Sequences of Random Vectors

James K Beard

Table of Contents

Introduction

In the presentation for the Philadelphia IEEE of November 16, 2021, an Appendix was provided that presented or expanded on several topics in the talk. One of these topics was on computing estimates of mean and covariance matrix from a sequence of random vectors. This is helpful when using particle filters, unscented transformations, and related methods of tracking and estimation. The presentation in that Appendix, while providing the correct expressions for the estimates, provides insufficient detail in the mathematics supporting these estimates. A simpler and more complete presentation of the mathematics of estimates of sample mean and covariance from a sequence of random vectors is given below.

Problem Statement

Classical Problem

We are presented with a sequence of N random vectors of K elements each, all with the same mean and with a Gaussian error with the same covariance, or at least with an error distributed according to a law that allows mean and variance to be defined. We wish to provide the most accurate possible estimate of the mean, and also the most accurate possible estimate of the covariance.

Sophisticated Tracker Problem

In some trackers, specifically particle filters and trackers based on unscented transforms, and other trackers based on a Bayesan estimate or the Chapman-Kolmogorov equation, the mean state vector is estimated by a weighted average of the noisy vectors as produced by modeling their values with specific values of the noise in the state vector propagation equations. The estimator of the covariance matrix must reflect this weighting.

Probability Functions

Gaussian, Zero Mean, Unity Variance

We begin with the probability density function of a Gaussian variable with zero mean and unity variance, from Abramowitz & Stegun 26.2.1:

$$p(x)=\frac{1}{\sqrt{2\pi}}\cdot\exp\left(-\frac{x^2}{2}\right) \quad . \tag{1}$$

We will take (1) as an axiom.

General Mean and Variance

If a variable y has mean μ and variance σ , we can write it as proportional to a zero mean, unity variance Gaussian random variable x ,

$$y=\sigma\cdot(x+\mu) \quad . \tag{2}$$

The probability density function of y becomes

$$p(y)=\frac{1}{\sqrt{2\pi\sigma}}\cdot\exp\left(-\frac{y^2}{2\sigma^2}\right) \quad . \tag{3}$$

Note that the probability density function has meaning as an integrand or the kernel of an integral, and the Jacobian determinant that follows from the differential of the integral with the variable change must be incorporated into the probability density function. In the case of (1), (2) and (3), the Jacobian determinant is the standard deviation, or the square root of the variance, σ .

Vector of Random Variables

The joint probability density function for K uncorrelated Gaussian random variables, all with zero mean zero and unity variance, is

$$p(\vec{y})=\prod_{i=1}^{K}\frac{1}{\sqrt{2\pi}}\cdot\exp\left(-\frac{y_i^2}{2}\right)=\frac{1}{\sqrt{(2\pi)^K}}\cdot\exp\left(-\frac{\vec{y}^T\cdot\vec{y}}{2}\right) \quad . \tag{4}$$

General Vector Mean and Covariance

We can use a vector of uncorrelated zero mean, unity variance random variables to produce a vector with a given mean $\vec{z}_0$ and covariance P with the variable change

$$\vec{z}=P^{1/2}\cdot\vec{y}+\vec{z}_0 \tag{5}$$

where $P^{1/2}$ is a matrix square root of P ,

$$\begin{array}{l} P=[P^{1/2}]^T\cdot[P^{1/2}] \text{ (upper trianguar Cholesky factor)} \\ \quad\text{or} \\ P=[P^{1/2}]\cdot[P^{1/2}]^T \text{ (lower triangular Cholesky factor)} \end{array} \quad . \tag{6}$$

The simplest way to obtain a matrix square root is by Cholesky decomposition. When P is positive definite, the Cholesky factorization is real, and is unique within the choice of sign for the K square roots in the decomposition. Other matrix square roots are related to either the upper triangular or lower triangular Cholesky factorization C by a similarity transformation,

$$P^{1/2}=M^{-1}\cdot C\cdot M \tag{7}$$

where M is any nonsingular $K\times K$ matrix.

Given (5), the probability density function of $\vec{z}$ is

$$p(\vec{z})=\frac{1}{\sqrt{(2\pi)^K|P|}}\cdot\exp\left(-\frac{1}{2}\cdot(\vec{z}-\vec{z}_0)^T\cdot P^{-1}\cdot(\vec{z}-\vec{z}_0)\right) \quad . \tag{8}$$

Note that we have included the Jacobian determinant of $|P|^{1/2}$. Equation (8) is given in Gelb, 2.2.39 page 34.

Set of N Random Vectors

A set of N random vectors with the same mean $\vec{z}_0$ and covariance P has the joint probability density function

$$\begin{aligned} p(\vec{z}_i,\ i=1\ldots N) &= \prod_{i=1}^{N}\frac{1}{\sqrt{(2\pi)^K|P|}}\cdot\exp\left(-\frac{1}{2}\cdot(\vec{z}_i-\vec{z}_0)^T\cdot P^{-1}\cdot(\vec{z}_i-\vec{z}_0)\right) \\ &= \frac{1}{\sqrt{(2\pi)^{K\cdot N}\cdot|P|^N}}\cdot\exp\left(-\frac{1}{2}\cdot\sum_{i=1}^{N}(\vec{z}_i-\vec{z}_0)^T\cdot P^{-1}\cdot(\vec{z}_i-\vec{z}_0)\right) \end{aligned} \quad . \tag{9}$$

Estimating Mean and Covariance

We use the method of maximum likelihood here, because of its simplicity and the fact that this approach has been shown to provide estimators that are unsurpassed in minimizing the variance of the estimators for any given data set.

The method of maximum likelihood requires that you write a Bayesian probability density function for the data $\vec{y}$, given true values $\vec{x}$ of the parameters to be estimated, which we will call the likelihood function $L(\vec{x})$:

$$L(\vec{x})=p(\vec{y}|\vec{x}) \tag{10}$$

By maximizing the likelihood function $L(\vec{x})$, the method finds the values of the parameters $\vec{x}$ which makes the observed data most likely.

Classical Unweighted Estimators

We use (9) to write the likelihood function

$$L(\vec{z}_0,P|\vec{z}_i)=\frac{1}{\sqrt{(2\pi)^{K\cdot N}\cdot|P|^N}}\cdot\exp\left(-\frac{1}{2}\cdot\sum_{i=1}^{N}(\vec{z}_i-\vec{z}_0)^T\cdot P^{-1}\cdot(\vec{z}_i-\vec{z}_0)\right) \quad . \tag{11}$$

This probability density function takes a simpler form when the natural logarithm is taken. Since the natural logarithm is monotonic, thus maximizing the peak not change the result, even though the equations are dramatically simplified. The log likelihood function is

$$l\left(\vec{z}_0,P\middle|\vec{z}_i\right)=-\frac{K\cdot N}{2}\cdot\ln\left(2\pi\right)-\frac{N}{2}\cdot\ln\left(|P|\right)-\frac{1}{2}\cdot\sum_{i=1}^{N}\left(\vec{z}_i-\vec{z}_0\right)^T\cdot P^{-1}\cdot\left(\vec{z}_i-\vec{z}_0\right)\quad . \tag{12}$$

We begin by maximizing the log likelihood function as a function of $\vec{z}_0$,

$$\frac{\partial l\left(\vec{z}_0,P\middle|\vec{z}_i\right)}{\partial\vec{z}_0}=\sum_{i=0}^{N}P^{-1}\cdot\left(\vec{z}_i-\vec{z}_0\right)=\vec{0} \tag{13}$$

which immediately leads us to the classical result, the sample mean:

$$\text{Est}\left(\vec{z}_0\right)=\frac{1}{N}\cdot\sum_{i=1}^{N}\vec{z}_i\quad . \tag{14}$$

Equation (14) is linear in the data, and such MLE-derived estimators are known to meet the Cramer-Rao bound, which is the minimum possible variance for the given data set. This minimum possible variance can be found from the second derivative of the log likelihood function, which produces the negative of the Fisher information matrix, and the Fisher information matrix is the inverse of the optimum covariance matrix:

$$-\text{Cov}\left\{\text{Est}\left(\vec{z_0}\right)\right\}^{-1}=\frac{\partial^2 l\left(\vec{z_0},P\middle|\vec{z_i}\right)}{\partial\vec{z_0^2}}=-N\cdot P^{-1} \tag{15}$$

or, the Cramer-Rao bound, which is achieved by the estimator given by (14), is

$$\text{Cov}\left\{\text{Est}\left(\vec{z}_0\right)\right\}=\frac{1}{N}\cdot P\quad . \tag{16}$$

Equation (16) gives the Cramer-Rao bound, and direct evaluation of the covariance of the estimate from (14) shows that this is the covariance of the estimate.

Maximizing the likelihood function as a function of P is a bit more complex. We note that derivatives of the trace or determinant of a matrix with respect to the matrix are defined classically (Gelb, 2.1-72 and 2.1-75 p. 23),

$$\begin{aligned}\frac{\partial}{\partial A}\text{trace}\left[B\cdot A\cdot C\right] &= B^T\cdot C^T\\ \frac{\partial}{\partial A}|B\cdot A\cdot C| &= |B\cdot A\cdot C|\cdot A^{-T}\end{aligned} \tag{17}$$

so that the second term in (12) is amenable to a derivative with respect to P and the third term is also, if we consider the scalar quadratic form to be the trace of a 1×1 matrix.

We can simplify the algebra by taking the partial derivatives with respect to P^{-1} rather than P , which has the effect of avoiding a multiplier of P^{-2} but we will take the partial derivative with respect to P so that we may more directly address the Cramer-Rao bound for $\mathrm{Est}(P)$ later. Taking the gradient of the log likelihood function (12) with respect to P and using (17),

$$\frac{\partial l(\vec{z}_0, P|\vec{z}_i)}{\partial P} = -\frac{N}{2} \cdot P^{-1} + \left[\frac{1}{2} \cdot \sum_{i=1}^{N} (\vec{z}_i - \vec{z}_0) \cdot (\vec{z}_i - \vec{z}_0)^T\right] \cdot P^{-2} = 0 \tag{18}$$

which gives us

$$\mathrm{Est}_{TD}(P) = \frac{1}{N} \cdot \sum_{i=1}^{N} (\vec{z}_i - \vec{z}_0) \cdot (\vec{z}_i - \vec{z}_0)^T \quad . \tag{19}$$

The issue that must be resolved to provide a practical estimator is that “truth” data for the mean, $\vec{x}_0$, is required to evaluate the estimator, and, of course, $\vec{x}_0$ is not available in practice; we indicate this problem with the subscript “TD.” Conventionally, the estimator of $\vec{x}_0$ given in (14) is used in lieu of truth data which introduces a magnitude bias, because $\vec{x}_0$ is correlated with each $\vec{x}_i$. Removing this bias gives us the classical result

$$\mathrm{Est}(P) = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (\vec{z}_i - \mathrm{Est}(\vec{z}_0)) \cdot (\vec{z}_i - \mathrm{Est}(\vec{z}_0))^T \quad . \tag{20}$$

Equation2 (19) and (20) are not linear in the data and thus (20) provides an estimator that does not achieve the Cramer-Rao bound. We do know from the properties of maximum likelihood estimators that the modified estimator of (20) is nonlinear, but the covariance of the estimator approaches the Cramer-Rao bound as N increases, and that deviation of the covariance from the Cramer-Rao bound is of order $1/N$.

The Cramer-Rao bound is found as the second gradient of the log likelihood function with respect to P , which is a result of tensor order 4; vectors are of tensor order 1 and matrices are of tensor order 2, and obtaining a simple expression for the covariance of P requires inverting a quantity of tensor order 4. Finding the Cramer-Rao bound for P is most simply treated by augmenting the set of N vectors of K elements each to form a single vector with K^2 elements, with a $K^2 \times K^2$ covariance matrix.

Within ordinary linear algebra principles, the elements of the $K^2 \times K^2$ covariance matrix can be computed in blocks of $K \times K$ matrices using conventional linear algebra methods as follows.

We write the gradient of P with respect to one of its elements $p_{i0,j0}$ by $P'_{i0,j0}$,

$$P'_{i0,j0} = \frac{\partial P}{\partial p_{i0,j0}} = [\delta(i - i0) \cdot \delta(j - j0)] \quad . \tag{21}$$

We find the gradient of P^{-1} with respect to an element of P from

$$\begin{aligned} P\cdot P^{-1} &= I \\ P'_{i0,j0}\cdot P^{-1}+P\cdot\frac{\partial P^{-1}}{\partial p_{i0,j0}} &= 0 \end{aligned} \qquad (22)$$

as

$$\frac{\partial P^{P-1}}{\partial p_{i0,j0}}=-P^{-1}\cdot P'_{i0,j0}\cdot P^{-1}=-P^{-1}\cdot[\delta(i-i0)\cdot\delta(j-j0)]\cdot P^{-1}=-\overrightarrow{pinv}_{i0}\cdot\overrightarrow{pinv}_{j0}^{T} \qquad (23)$$

where $\overrightarrow{pinv}_i$ is column i of P^{-1} (which is the same as row i of P^{-1}).

A total of K^2 elements of the $K^2\times K^2$ Fisher information matrix is given by the elements of

$$\begin{aligned} \frac{\partial^2 l(\vec{z}_0,P|\vec{z}_i)}{\partial P\cdot\partial p_{i0,j0}} &= +\frac{N}{2}\cdot\overrightarrow{pinv}_{i0}\cdot\overrightarrow{pinv}_{j0}^{T} \\ &\quad -\left[\sum_{i=1}^{N}(\vec{z}_i-\vec{z}_0)\cdot(\vec{z}_i-\vec{z}_0)^T\right]\cdot\left[P^{-1}\cdot\overrightarrow{pinv}_{i0}\cdot\overrightarrow{pinv}_{j0}^{T}+\overrightarrow{pinv}_{i0}\cdot\overrightarrow{pinv}_{j0}^{T}\cdot P^{-1}\right] \end{aligned} \quad . \qquad (24)$$

Estimate of Vector by Weighted Average

The particle Kalman filter and the unscented Kalman filter use a weighted average to estimate the state vector,

$$\mathrm{Est}_W(\vec{x}_0)=\frac{\sum_{i=1}^{N} w_i\cdot\vec{x}_i}{\sum_{i=1}^{N} w_i} \quad . \qquad (25)$$

The covariance of this estimate is found directly as

$$\mathrm{Cov}\{\mathrm{Est}_W(\vec{x}_0)-\vec{x}_0\}=\frac{\sum_{i=1}^{N} w_i^2}{\left(\sum_{i=1}^{N} w_i\right)^2}\cdot P \quad . \qquad (26)$$

An estimate of this covariance is available by using the estimate of P from (20). Note that, for nonnegative w_i ,

$$\frac{1}{N}\leq\frac{\sum_{i=1}^{N} w_i^2}{\left(\sum_{i=1}^{N} w_i\right)^2}\leq 1 \quad . \qquad (27)$$

Equality to $1/N$ is achieved with all of the w_i are equal, which is the unweighted case. Equality to 1 is achieved when only one of the weights is nonzero.

References

[DLMF] NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.1.3 of 2021-09-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds. (Hardback with CD-ROM ISBN-13: 9780521192255; Paperback with CD-ROM ISBN-13: 9780521140638), also available online as referenceable HTML at https://dlmf.nist.gov/

Handbook of Mathematical Functions, M. Abramowitz & I. Stegun, Eds., National Bureau of Standards Applied Mathematics Series, #55 (1972) (Dover paperback ISBN 978-0486612720).

Applied Optimal Estimation, A. Gelb, Ed., MIT Press (1974) ISBN 978-0262570480.

Keywords for literature searches: particle [Kalman] filter, unscented [Kalman] filter, Mahalanobis Distance, Chapman-Kolmogorov equation, statistical efficiency, Fisher information matrix, importance sampling, maximum likelihood