# Concentration Ellipsoids

## ECE275A – Lecture Supplement – Fall 2008

Kenneth Kreutz–Delgado
Electrical and Computer Engineering
Jacobs School of Engineering
University of California, San Diego

VERSION LSECE275CE–F08v1.0



Let $\tilde{\theta} = \hat{\theta} - \theta$ denote the error associated with a uniformly unbiased estimator[1] $\hat{\theta}$ of an $n$-dimensional real parameter vector $\theta$. Necessarily then, $\mathrm{E}_\theta\left\{\tilde{\theta}\right\} = 0$. Let the associated real $n \times n$ error covariance matrix of $\tilde{\theta}$ be given by

$$\Sigma = \Sigma(\hat{\theta}, \theta) = \mathrm{Cov}_\theta\left\{\tilde{\theta}\right\} = \mathrm{E}_\theta\left\{\tilde{\theta}\tilde{\theta}^T\right\} = \mathrm{E}_\theta\left\{(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T\right\} .$$

Note that $\Sigma(\hat{\theta}, \theta)$ is a function of both the estimator $\hat{\theta}$ and the assumed (unknown) parameter vector $\theta$, even though we often use the simpler notation $\Sigma$ in place of the more informative $\Sigma(\hat{\theta}, \theta)$, and thereby not show this dependence explicitly.[2] It is usual to further assume that the (necessarily symmetric and positive-semidefinite) error covariance matrix is positive-definite (and hence full-rank and invertible), $\Sigma > 0$. In the latter case the random vector $\tilde{\theta}$ is called *full-rank*.

It is well-known that a symmetric $n \times n$ positive-definite matrix, $\Sigma$, has $n$ nonzero real eigenvalues, $\sigma_i > 0$, and associated real orthogonal (which can assumed to be normalized) eigenvectors $u_i$,

$$\Sigma\, u_i = \sigma_i\, u_i \quad \text{with} \quad \langle u_i, u_j \rangle = u_i^T u_j = \delta_{ij} \quad \forall\, i, j = 1, \cdots, n\,. \tag{1}$$

Defining the real $n \times n$ orthogonal matrix $U = [u_1 \cdots u_n]$, and the real diagonal matrix $\Lambda = \mathrm{diag}\,[\sigma_1 \cdots \sigma_n]$ we can rewrite (1) as

$$\Sigma = U \Lambda U^T \quad \text{with} \quad U^{-1} = U^T\,. \tag{2}$$

[1] Also known as an absolutely unbiased, or (in the Bayesian context) a conditionally unbiased estimator.

[2] In lecture, at various times we also use the notations $\Sigma_\theta(\hat{\theta})$ (both dependencies explicit), $\Sigma(\hat{\theta})$ (the $\theta$ dependence suppressed), $\Sigma_\theta$ (the $\hat{\theta}$ dependence suppressed), and $\Sigma_{\hat{\theta}}$ (the $\theta$ dependence suppressed).

From (2) one can easily show that $\Sigma^k = U\Lambda^k U^T$ for *all integer* (positive and negative) values of $k$, and meaningfully generalize this fact by *defining* $\Sigma^r$ for *all real* values of $r$ to be $\Sigma^r = U\Lambda^r U^T$, where $\Lambda^r = \mathrm{diag}\,[(\sigma_1)^r \cdots (\sigma_n)^r]$. Note that using this particular definition results in $\Sigma^r$ being symmetric positive definite for all $r$. In particular, we have $\Sigma^{-1} = U\Lambda^{-1}U^T$ and $\Sigma^{\frac{1}{2}} = U\Lambda^{\frac{1}{2}}U^T$.

The representation $\Sigma = U\Lambda U^T$ can be equivalently written as

$$\Sigma = \sigma_1 u_1 u_1^T + \cdots + \sigma_n u_n u_n^T \,. \tag{3}$$

The representation (3) is a particular example of a so-called rank–one expansion of a matrix,[3] which, because it is an expansion in terms of the eigenpairs $(\sigma_i, u_i)$, $i = 1, \cdots, n$, is known as the *Spectral Representation* of the positive definite matrix $\Sigma$. It is also the singular value decomposition (SVD) of $\Sigma$. Because $\Sigma$ is symmetric positive–definite, its singular values and eigenvalues happen to be identical, but more generally this is not the case. The SVD is a generalization of the Spectral Representation that can be applied to general non–symmetric, non–diagonalizable and non–square matrices.

Because the symmetric positive-definite matrix $\Sigma$ happens to also be a covariance matrix, the Spectral Representation (3) is also known as a *Karhunen–Loéve expansion* (or K-L expansion)[4] and as a *Principal Components Analysis* expansion (or PCA expansion). In the latter case the eigenpairs $(\sigma_i, u_i)$ are known as the *principal Components*. The unit vectors $u_i$ are known as the principal directions (in $\tilde{\theta}$–space).

[3] Each individual term $u_i u_i^T$ is a rank-one matrix, as can be easily shown. A more general rank–one expansion of an *arbitrary*, and even possibly *non-square*, matrix which we have previously encountered is the SVD. Indeed, equations (2)-(3) give the SVD of the special, positive-definite matrix $\Sigma$, and $\sigma_i$, $i = 1, \cdots n$ are also the singular values of $\Sigma$. It is to emphasize this fact that I use $\sigma_i$ to denote the eigenvalues (which, in this case, are also the singular values) of $\Sigma$ rather than the (perhaps) more conventional notation of $\lambda_i$.

[4] The K-L expansion provides a representation of the components of the random vector $\tilde{\theta}$, viewed as a correlated stochastic time series (also known as a *random process*) $\tilde{\theta}[1], \cdots, \tilde{\theta}[n]$, in terms of the components of the nonrandom vector $u_i$ viewed as a nonrandom time-series $u_i[1], \cdots, u_i[n]$. The $n$ deterministic processes $u_i[k]$, $i = 1, \cdots, n$, are orthonormal with respect to the $\ell_2$ inner product $\langle u, v\rangle = \sum_{k=1}^{n} u[k]v[k]$. The K-L expansion (or representation) is given by $\tilde{\theta}[k] = \sum_{i=1}^{n} \tilde{\pi}[i]u_i[k]$, where $\tilde{\pi}[i] = \left\langle u_i, \tilde{\theta}\right\rangle = u_i^T\tilde{\theta}$ is random. If we define the vector $\tilde{\pi} = [\tilde{\pi}[1], \cdots, \tilde{\pi}[n]]^T$, we can readily show that $\tilde{\pi} = U^T\tilde{\theta}$, $\tilde{\theta} = U\tilde{\pi}$, and that $\mathrm{E}_\theta\,(\tilde{\pi}\tilde{\pi}^T) = \Lambda$. This shows that we can transform the correlated discrete-time random process $\tilde{\theta}[k]$ into an equivalent uncorrelated discrete-time random process $\tilde{\pi}[k]$.

More generally (and the usual situation encountered in communications theory) The K–L expansion allows a *continuous-time* random process $\tilde{\theta}(t)$ to be represented in terms of a countable set of continuous-time orthogonal (with respect to the $\mathcal{L}_2$ inner product) deterministic functions $u_i(t)$, $\tilde{\theta}(t) = \sum_{i=1}^{\infty} \tilde{\pi}[i]u_i(t)$, where $\tilde{\pi}[i]$ is uncorrelated. This results in the transformation of the *continuous-time correlated* random process $\tilde{\theta}(t)$ into an equivalent *discrete-time uncorrelated* random process $\tilde{\pi}[k]$, $k = 1, 2, \cdots$, which can be quite useful in a theoretical anaysis. For example, this transformation is commonly done when analyzing intersymbol interference (ISI) in a communication channel.

The principal components are not random. However, the projections of the random vector $\tilde{\theta}$ onto the nonrandom principal directions $u_i$, denoted by $\tilde{\pi}[i] = u_i^T \tilde{\theta}$, $i = 1, \cdots, n$, results in a collection of uncorrelated scalar random variables each with variance $\sigma_i$,

$$\mathrm{E}_\theta \left\{ \tilde{\pi}[i]\, \tilde{\pi}[j] \right\} = \sigma_i \delta_{ij} \,.$$

If we further define the random vector

$$\tilde{\pi} = [\tilde{\pi}[1], \cdots, \tilde{\pi}[n]]^T = U^T \tilde{\theta}$$

this corresponds to

$$\mathrm{E}_\theta \left\{ \tilde{\pi} \tilde{\pi}^T \right\} = \Lambda \,.$$

Note that it is readily shown that,

$$\mathrm{mse}_\theta(\hat{\theta}) = \mathrm{E}_\theta \left\{ \|\tilde{\theta}\|^2 \right\} = \mathrm{trace}\, \Sigma = \mathrm{trace}\, \Lambda = \mathrm{E}_\theta \left\{ \|\tilde{\pi}\|^2 \right\} = \sigma_1 + \cdots + \sigma_n \,.$$

Assuming that the eigenvalues are ordered from largest to smallest as $\sigma_1 \geq \cdots \geq \sigma_n > 0$ (as is usually done), then the corresponding principal directions $u_1, \cdots, u_n$, describe directions in $\tilde{\theta}$–space which explain the observed statistical variation of $\tilde{\theta}$ in decreasing importance as measured by the eigenvalues. The projection, $\tilde{\pi}[i]$, of $\tilde{\theta}$ onto $u_i$ is known as the loading of $\tilde{\theta}$ on the principal direction $u_i$, and it explains $100 \times \frac{\sigma_i}{\sigma_1 + \cdots + \sigma_n}$ % of the observed statistical variation in $\tilde{\theta}$. Sometimes a principal direction can be physically (or otherwise) interpreted, thereby providing an explanation for the amount of variation seen along that particular direction. In terms of $\tilde{\pi}$, we can recover $\tilde{\theta}$ as

$$\tilde{\theta} = U U^T \tilde{\theta} = U \tilde{\pi} = \tilde{\pi}[1] u_1 + \cdots + \tilde{\pi}[n] u_n \,.$$

A so-called rank-$r$ approximation to $\tilde{\theta}$ in terms of its $r < n$ first most important principal components as measured in order of decreasing importance by $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq \cdots \geq \sigma_n$, is provided by

$$\tilde{\theta}^{(r)} = \tilde{\pi}[1] u_1 + \tilde{\pi}[2] u_2 + \cdots + \tilde{\pi}[r] u_r \,.$$

Note that

$$\mathrm{trace}\ \mathrm{Cov}_\theta \left\{ \tilde{\theta}^{(r)} \right\} = \mathrm{trace}\ \mathrm{E}_\theta \left\{ \tilde{\theta}^{(r)} \tilde{\theta^{(r)}}^T \right\} = \sigma_1 + \cdots + \sigma_r \,,$$

and

$$\mathrm{trace}\ \mathrm{E}_\theta \left\{ \left( \tilde{\theta} - \tilde{\theta}^{(r)} \right) \left( \tilde{\theta} - \tilde{\theta}^{(r)} \right)^T \right\} = \sigma_{r+1} + \cdots + \sigma_n \,.$$

The rank-$r$ approximation $\tilde{\theta}^{(r)}$ explains $100 \times \frac{\sigma_1 + \cdots + \sigma_r}{\sigma_1 + \cdots + \sigma_r + \cdots + \sigma_n}$ % of the observed variation in $\tilde{\theta}$. Finally, if the components of $\tilde{\theta}$ are interpreted as a correlated random sequence, then the components of $\tilde{\pi}$ can be interpreted as an uncorrelated ("whitened") random sequence which is entirely equivalent to $\tilde{\theta}$ (see also footnote 4).

## Homework Problems.

The symmetric positive-definite matrices $\Sigma_1$ and $\Sigma_2$ denote the error covariances associated with the unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ respectively. The respective estimation errors are denoted by $\tilde{\theta}_1$ and $\tilde{\theta}_2$. It is convenient to make the transformation $\lambda_i^2 = \sigma_i$, $i = 1, \cdots, n$, so that $\lambda_i = \sqrt{\sigma_i}$ denotes the *standard deviation* of the statistical variance of $\tilde{\theta}$ along the direction $u_i$.

1. Show that a level surface of $\tilde{\theta}^T \Sigma^{-1} \tilde{\theta} = k^2$ is the surface of an $n$-dimensional hyperellipsoid in $\tilde{\theta}$–space (parameter error space) and give the semi-major axes of the hyperellipsoid. Describe the region $\tilde{\theta}^T \Sigma^{-1} \tilde{\theta} \leq k^2$. Hint: Transform $\tilde{\theta}$ into the uncorrelated-elements random vector $\tilde{\pi} = U^T \tilde{\theta}$ (see the discussion in footnote 4) and determine the form of the level surfaces in the $\tilde{\pi}$–space.

2. Let $0 \leq \Sigma_1 \leq \Sigma_2$. Let $a$ be an arbitrary unit vector in $\tilde{\theta}$–space. (i) Show that the variance of $a^T \tilde{\theta}_1$ is less than or equal to the variance of $a^T \tilde{\theta}_2$ for any direction $a$ and give an interpretation of this fact. (ii) For an arbitrary fixed direction $a$ consider events of the form $\left\{ \left| a^T \tilde{\theta}_i \right| \leq L_i \right\}$, $0 \leq L_i < \infty$, $i = 1, 2$, each having a probability of occurrence of at least $p$,

$$P\left( \left| a^T \tilde{\theta}_i \right| \leq L_i \right) \geq p > 0, \qquad i = 1, 2\,.$$

   Use the Chebyshev inequality[5]

$$P\left( |e| \leq L \right) \geq 1 - \frac{\mathrm{E}\left\{ e^2 \right\}}{L^2}\,.$$

   to determine values $L_1$ and $L_2$ satisfying the condition $L_1 \leq L_2$ and give a mathematical relationship between these two bounds. Interpret this result.

3. Prove that if $0 < \Sigma_1 \leq \Sigma_2$, then $\Sigma_1^{-1} \geq \Sigma_2^{-1} > 0$. This result, which is needed to answer the next question below, can be surprisingly hard to prove. If you can't fully prove it, just show some of your attempts and move on to the next problem.[6]

4. Now consider events of the form $\left\{ \tilde{\theta}_i^T \Sigma_i^{-1} \tilde{\theta}_i \leq k^2 \right\}$, for $0 \leq k < \infty$, $i = 1, 2$.

   (i) Explain that such a region defines a hyperellipsoid volume in $\tilde{\theta}$–space (parameter error space) which is associated with the estimation error for the estimator $\hat{\theta}_i$. (ii) For a given probability of occurrence of the events of at least $P$, use the Tchebycheff inequality to determine a value for $k$. (iii) For the same fixed value of $k$, show that the

[5] Because of ambiguous transliteration from the Cyrillic (Russian) alphabet, there are a variety of equivalent spellings. E.g., "Tchebycheff," "Chebyshev," "Tchebychev," etc. The Chebyshev inequality is an important statistical result which can be found in the standard upper division electrical engineering probability and statistics textbooks. The form of the Chebyshev inequality used here is one of several equivalent variant statements.

[6] Hint: Show that the result holds for *diagonal* covariance matrices. Then show that you can *simultaneously diagonalize* two arbitrary covariance matrices and use your result for diagonal matrices.

hyperellipsoid volume associated with $\tilde{\theta}_1$ is completely contained in the hyperellipsoid volume associated with $\tilde{\theta}_2$. (iv) Explain how the partial ordering $\Sigma_1 \leq \Sigma_2$ can be interpreted in terms of nested "concentration ellipsoid" volumes[7] and, in a planar drawing, symbolically represent the ellipsoidal volumes of two matrices that cannot be ordered with respect to each other.

**Comment on the homework problems.** Not surprisingly, more precise statements can be made under a multivariate gaussian assumption on the error $\tilde{\theta}$. See, for example, the discussion on pages 225–226 in *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis,* by L.L. Scharf (1991, Addison-Wesley).

[7] That is, where is the estimation error, with probability $P$, likely to be be concentrated.