2011

Problems in GPS Accuracy

Michael Thomas Vodhanel
Claremont Graduate University

Recommended Citation
Vodhanel, Michael Thomas, "Problems in GPS Accuracy" (2011). *CGU Theses & Dissertations.* Paper 22.
http://scholarship.claremont.edu/cgu_etd/22

DOI: 10.5642/cguetd/22

Problems in GPS Accuracy

by

Michael Thomas Vodhanel

A dissertation submitted to the Faculty of Claremont Graduate University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
(Applied Mathematics)

Claremont Graduate University

2011

This dissertation has been duly read, reviewed, and critiqued by the Committee listed below, which hereby approves the manuscript of Michael Thomas Vodhanel as fulfilling the scope and quality requirements for meriting the degree of Doctor of Philosophy.

John E. Angus, Chair

Claremont Graduate University

School of Mathematical Sciences

Professor

Allon Percus

Claremont Graduate University

School of Mathematical Sciences

Associate Professor

Ali Nadim

Claremont Graduate University

School of Mathematical Sciences

Professor

Abstract of the Dissertation

Problems in GPS Accuracy

by

Michael Thomas Vodhanel

Claremont Graduate University: 2011

Improving and predicting the accuracy of positioning estimates derived from the global positioning system (GPS) continues to be a problem of great interest. Dependable and accurate positioning is especially important for navigation applications such as the landing of commercial aircraft. This subject gives rise to many interesting and challenging mathematical problems. This dissertation investigates two such problems.

The first problem involves the study of the relationship between positioning accuracy and satellite geometry configurations relative to a user's position. In this work, accuracy is measured by so-called dilution of precision (DOP) terms. The DOP terms arise from the linear regression model used to estimate user position from GPS observables, and are directly related to user position errors. An analysis of the statistical properties explaining the behavior of the DOP terms is presented. The most accurate satellite geometries and worst configurations are given for some cases.

The second problem involves finding methods for detecting and repairing cycle-slips in range delay data between a satellite and a receiver. The distance between a satellite and a receiver can be estimated by measuring the difference in the carrier frequency phase shift experienced between the satellite and receiver oscillators. Cycle-slips are discontinuities in the integer number of complete cycles in these data, and are caused by interruptions or degradations in the signal such as low signal to noise ratio, software failures, or physical obstruction of the signals. These slips propagate to errors in user positioning. Cycle-slip detection and repair are crucial to maintaining accurate positioning. Linear regression models and sequential hypothesis testing are used to model, detect, and repair cycle-slips. The effectiveness of these methods is studied using data obtained from ground-station receivers.

Acknowledgments

First and foremost I would like to thank my advisor Dr. John Angus. He continuously offered insight that helped me to improve my work and devoted a very generous amount of time to reviewing both my mathematics and writing. His knowledge of the GPS system was of great benefit to me and I believe that I achieved a much higher standard as a direct result of his input throughout this process.

I would also like to thank my committee members Dr. Ali Nadim and Dr. Allon Percus. They both offered great advice that helped me to create a higher quality work. Their experience was invaluable in helping me to avoid pitfalls that I would not have recognized otherwise.

I would like to show very special appreciation to Dr. Eric Altshuler of Sequoia Research Corporation who originally introduced me to the applications addressed in this dissertation. Much of my knowledge about the fundamentals of the GPS system comes from working with him and I have benefited from his expertise countless times in the last few years.

I would also like to thank my parents who always worked hard and sacrificed so that I could get an excellent education and who have supported me through all of my years in school.

Finally, I would like to thank Dr. Hedley Morris, a wonderful teacher who showed enthusiasm at the prospect of helping me with this process but passed away before the time came. It is a great loss to his students.

I would not have been able to complete this dissertation without the help and guidance of the people mentioned here.

Table of Contents

Glossary of Terms

ANOVA - Analysis of Variance. A method of creating statistical hypothesis tests in which the total sum of squares is broken into components which have meaning with regards to the models' fit of the data, such as the sum of squares regression and sum of squares residuals.

BLUE - Best Linear Unbiased Estimator. A parameter estimator which has least variance of all unbiased estimators that are linear functions of the observables.

CSR - Cycle-slip Repair Tool. The software tool developed for the purpose of detecting and repairing cycle-slips in RINEX files.

DOP - Dilution of Precision. A term proportional to the standard deviation of linear regression estimators and a widely used metric for user positioning accuracy.

Galileo - European global navigation satellite system.

GLONASS - Global Orbiting Navigation Satellite System (Russian global navigation satellite system).

GNSS - Global navigation satellite system. Such systems include GPS, Galileo, and GLONASS.

GPS - The Global Positioning System (United States global navigation satellite system).

IFPR - iono-free pseudorange. A pseudorange measurement free of ionospheric delay.

PRN - Pseudo-Random Noise. A generated binary signal with noise-like properties used as an identification code for a GPS satellite.

Pseudorange - An approximation of the range between a satellite and receiver which contains some error due to a bias in the receiver clock.

RINEX - The receiver independent exchange format. A widely used format for files which contain carrier and code data. See (Gurtner & Estey, 2007) for a detailed description.

WAAS - Wide Area Augmentation System. Augmentation of GPS aimed at improving accuracy, availability, and integrity.

WRS - WAAS Reference Station. Ground stations that monitor GPS satellite data.

1.1 Introduction to Dilution of Precision

A great deal of effort has been undertaken to quantify the accuracy of navigation systems that make use of the Global Positioning System (GPS). A good example of such a system is the Wide Area Augmentation System (WAAS), which is now used extensively by commercial airlines. Dilution of Precision (DOP), which is dependent on satellite positioning relative to the user, is a widely used metric for accuracy because it provides a proportionality constant between the precision of satellite range measurement and user position error (Misra & Enge, 2006). However, due to a formulation which is very difficult to conceptualize, the behavior of DOP with respect to user positioning has not been well understood. As a result, optimal satellite configurations which minimize the various DOPs have remained unknown. In section 1, a method for visualizing the behavior of DOP using parallelotopes is developed along with results needed to solve the optimization problems. Optimal satellite configurations are presented for various DOPs using this understanding and worst case satellite configurations are also discussed. Finally, using this information, an algorithm is developed for fast satellite subset selection which is superior to previously used methods. This is an important problem, as there are typically more satellites visible than channels available to a GPS receiver to track them, so it is advantageous for the receiver to fill its available channels with the combination of satellites that yields the most accurate navigation information.

1.2 Development of DOP

The purpose of this section is to introduce the usual formulation for DOP in both the general case and specifically for GPS. This section describes the basics of linear regression and GPS satellite geometry needed to understand DOP. Sections 1.3-1.5 establish a thorough understanding of the behavior of DOP as it is formulated in this section.

1.2.1 Basics of Linear Regression

Linear regression is a method of modeling the relationship between an observable y and a set of input variables $x_1,\ldots,x_n$ as a linear function:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \qquad \text{Eq. 1.2.1-1}$$

where ε is a random variable representing noise and each β_j is an unknown model parameter. By collecting m observations for various input values, m equations are formed. For $1 \le i \le m$:

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_n x_{i,n} + \varepsilon_i \qquad \text{Eq. 1.2.1-2}$$

These m equations can be written in matrix form.

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon} \qquad \text{Eq. 1.2.1-3}$$

Here, $\vec{y}$ is a column vector of m observable values, X is an m by n matrix of input values called the design matrix, $\vec{\beta}$ is a column vector of the unknown model parameters, and $\vec{\varepsilon}$ is a column vector of m random noise terms. Typically, the goal of this formulation is to find an estimate for the unknown model parameters $\vec{\beta}$. It will be assumed throughout that the matrix X is of full rank which implies that $X^T X$ is non-singular. In this case, the linear regression estimator for $\vec{\beta}$ is given by

$$\vec{b} = (X^T X)^{-1} X^T y \qquad \text{Eq. 1.2.1-4}$$

This linear regression estimator is the best linear unbiased estimator (BLUE) for $\vec{\beta}$ when the noise terms are independent and identically distributed with a finite common variance and zero mean. It is unbiased because the expected value is $E(\vec{b}) = \vec{\beta}$. It is best in the sense that out of all unbiased estimators that are linear functions of the observables $\vec{y}$, each b_i has the minimum variance. Under the assumption that each ε_i is independent and identically distributed with variance σ^2, the covariance of the estimation vector is given by:

$$\operatorname{cov}(\vec{b}) = \sigma^2 (X^T X)^{-1} \qquad \text{Eq. 1.2.1-5}$$

(Myers & Milton, 1998).

1.2.2 Basics of GPS Positioning

The GPS system currently consists of 31 satellites. Each continually transmits a signal that includes the time that the message was sent and the satellite position. From this information, pseudorange equations are formed for a user who has unknown position.

$$\rho_i(x, y, z, \Delta t) = \sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2} + c\Delta t \qquad \text{Eq. 1.2.2-1}$$

Here there are m equations, one for each satellite in view of the user $1 \le i \le m$. Each (x_i, y_i, z_i) is a known position for the i-th satellite. (x, y, z) is the unknown position of the user. Δt is a time offset of the user's GPS receiver from the GPS system time and c is the speed of light, which is the speed at which the signals propagate from the satellite to the user (Kaplan & Hegarty, 2006). The GPS receiver provides a noisy measurement of $\rho_i(x, y, z, \Delta t)$, namely $\rho_i(x, y, z, \Delta t) + \varepsilon_i$, where ε_i represents the random error term in the measurement.

One way to apply the aforementioned linear statistical model to solve these equations for the user position is to linearize them about a known, approximate user position $(\tilde{x}, \tilde{y}, \tilde{z})$ such that

$$(x, y, z) = (\tilde{x}, \tilde{y}, \tilde{z}) + (\Delta x, \Delta y, \Delta z) \qquad \text{Eq. 1.2.2-2}$$

A pseudorange equation can be linearized about $(\tilde{x}, \tilde{y}, \tilde{z})$ using Taylor expansion:

$$\rho_i(x, y, z, \Delta t) - \sqrt{(x_i - \tilde{x})^2 + (y_i - \tilde{y})^2 + (z_i - \tilde{z})^2} = \frac{\partial \rho_i(\tilde{x}, \tilde{y}, \tilde{z}, \Delta t)}{\partial \tilde{x}} \Delta x + \frac{\partial \rho_i(\tilde{x}, \tilde{y}, \tilde{z}, \Delta t)}{\partial \tilde{y}} \Delta y + \frac{\partial \rho_i(\tilde{x}, \tilde{y}, \tilde{z}, \Delta t)}{\partial \tilde{z}} \Delta z + c\Delta t \qquad \text{Eq. 1.2.2-3}$$

so that approximately, we have

$$-\rho_i(x,y,z,\Delta t)+r_i=\frac{x_i-\tilde{x}}{r_i}\Delta x+\frac{y_i-\tilde{y}}{r_i}\Delta y+\frac{z_i-\tilde{z}}{r_i}\Delta z-c\Delta t \qquad \text{Eq. 1.2.2-4}$$

where $r_i=\sqrt{(x_i-\tilde{x})^2+(y_i-\tilde{y})^2+(z_i-\tilde{z})^2}$. Note that the first three terms on the right hand side of this equation make up the dot product of the unit direction vector pointing from the linearization point to the satellite with the unknown offset of the actual user position from the linearization point. Taking our observables to be $y_i=-(\rho_i(x,y,z,t)+\varepsilon_i)+r_i$, and denoting:

$$\vec{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad \vec{\beta}=\begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \\ -c\Delta t \end{bmatrix}, \quad \vec{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix}$$

$$G=\begin{bmatrix} \frac{x_1-\tilde{x}}{r_1} & \frac{y_1-\tilde{y}}{r_1} & \frac{z_1-\tilde{z}}{r_1} & 1 \\ \frac{x_2-\tilde{x}}{r_2} & \frac{y_2-\tilde{y}}{r_2} & \frac{z_2-\tilde{z}}{r_2} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ \frac{x_m-\tilde{x}}{r_m} & \frac{y_m-\tilde{y}}{r_m} & \frac{z_m-\tilde{z}}{r_m} & 1 \end{bmatrix} \qquad \text{Eq. 1.2.2-5}$$

this yields the matrix system $\vec{y}=G\vec{\beta}+\vec{\varepsilon}$. The incremental component $\vec{\beta}$ can be estimated using linear regression. The user position may be approached by iterating the approximate position using $(\tilde{x},\tilde{y},\tilde{z})+(\Delta x,\Delta y,\Delta z)\rightarrow(\tilde{x},\tilde{y},\tilde{z})$. In practice, the iteration converges rapidly.

1.2.3 The Geometry Matrix and DOP

In the matrix G given by equation 1.2.2-5, the position portion of the i-th row is a unit vector pointing from the approximate user location (i.e. the linearization point) to the i-th satellite. The local level coordinate system is chosen such that:

$$\begin{bmatrix} \dfrac{x_i - \tilde{x}}{r_i} & \dfrac{y_i - \tilde{y}}{r_i} & \dfrac{z_i - \tilde{z}}{r_i} \end{bmatrix} = \begin{bmatrix} e_i & n_i & u_i \end{bmatrix} \qquad \text{Eq. 1.2.3-1}$$

where e, n, and u are east, north and up components respectively. So, these are the normalized satellite coordinates in the local coordinate east-north-up (ENU) system with the user at the origin. Thus, the G matrix can be written

$$G = \begin{bmatrix} \vec{e} & \vec{n} & \vec{u} & \mathbf{1} \end{bmatrix} \qquad \text{Eq. 1.2.3-2}$$

with the bold **1** representing a column vector of all ones. A very common alternate representation using local azimuth and elevation is

$$G = \begin{bmatrix} \cos(el)\sin(az) & \cos(el)\cos(az) & \sin(el) & \mathbf{1} \end{bmatrix} \qquad \text{Eq. 1.2.3-3}$$

(Parkinson & Spilker, 1996; Misra & Enge, 2006). This is simply a parametric (spherical coordinates) form of equation 1.2.3-2. Because it holds the satellite position data relative to the user, G will also be called the geometry matrix.

In the linear regression model from the previous section, G is the design matrix. From equation 1.2.1-5, the covariance of the position solution is $\sigma^2 (G^T G)^{-1}$. So $(G^T G)^{-1}$ holds information about amplification of the variance onto the positional solutions (Kaplan &

Hegarty, 2006). This information is called dilution of precision (DOP). The DOP terms are defined by:

$$\left(G^T G\right)^{-1} = \begin{bmatrix} EDOP^2 & \bullet & \bullet & \bullet \\ \bullet & NDOP^2 & \bullet & \bullet \\ \bullet & \bullet & VDOP^2 & \bullet \\ \bullet & \bullet & \bullet & TDOP^2 \end{bmatrix} \qquad \text{Eq. 1.2.3-4}$$

(Kaplan & Hegarty, 2006; Misra & Enge, 2006). These define DOPs for the east, north, vertical (up), and time directions. Other DOP terms are derived as square roots of sums of these. The commonly considered ones are:

$$Horizontal: \quad HDOP = \sqrt{EDOP^2 + NDOP^2} \qquad \text{Eq. 1.2.3-5}$$

$$Positional: \quad PDOP = \sqrt{EDOP^2 + NDOP^2 + VDOP^2} \qquad \text{Eq. 1.2.3-6}$$

$$Geometric: \quad GDOP = \sqrt{PDOP^2 + TDOP^2} = \sqrt{trace((G^T G)^{-1})} \qquad \text{Eq. 1.2.3-7}$$

Each of these gives information about the accuracy of user positioning in the described direction.

1.2.4 Conceptualization of DOP

The commonly used illustration of dilution of precision with two satellites in 2D is presented in the figures below. Assuming that there are some error bounds on the range of each satellite to the user, the user is within a region of error.

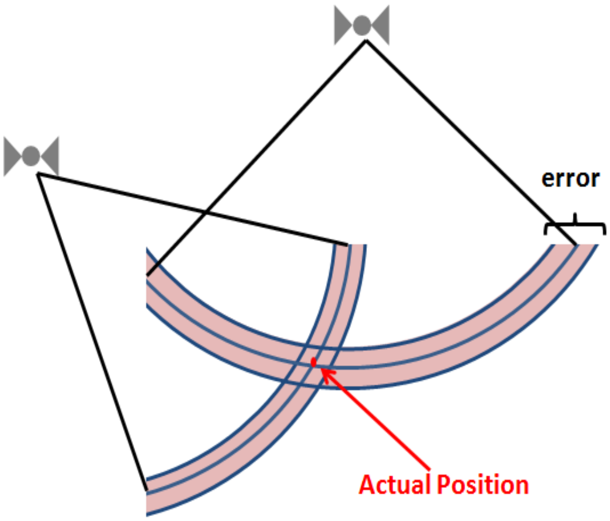


Figure 1.2.4-1

Figure 1.2.4-1 is a picture for a desirable DOP since the actual user position is within a relatively small region of error. With bad positioning, two satellites with the same range error can produce a much larger region of error.

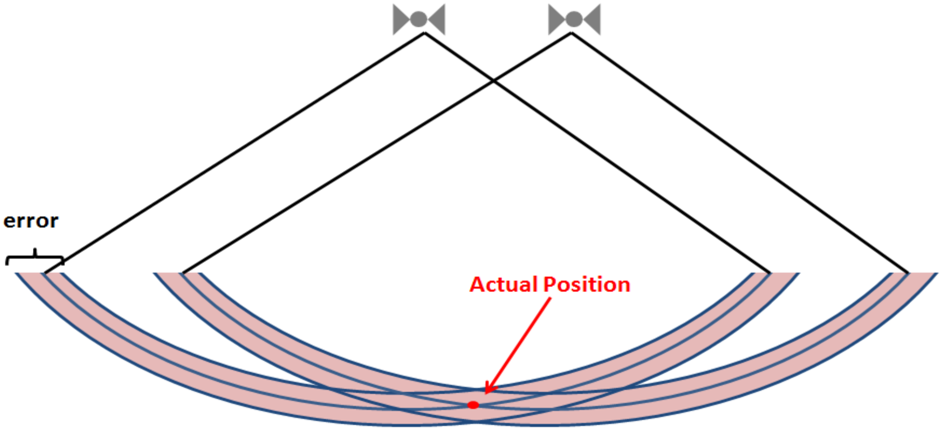


Figure 1.2.4-2

Figure 1.2.4-2 is an undesirable DOP. In the following sections, a much more thorough conceptualization of DOP behavior will be developed.

1.3 Parallelotope Theory for DOP

In this section, a method for understanding and visualizing the behavior of DOP is developed. DOP as it is presented in the definition, as a term in the inverse $G^T G$ matrix, is difficult to understand. It is a function of all the terms of the G matrix and thus has high dimensionality when m and n are not small. This section will present DOP as a function of only a few variables which represent an easy to understand geometric figure.

1.3.1 Properties of Gramian Matrices and their Submatrices

The first step in understanding the behavior of DOP is to observe that the covariance matrix of the linear regression system is the inverse of a Gramian matrix H .

Def (1.3.1-1): An n by n matrix H is Gramian if there exists some set of column vectors $v_1, \ldots, v_n$ such that $H_{ij} = v_i^T v_j$, $\forall_{1 \le i, j \le n}$.

It follows directly from the definition that any Gramian matrix is Hermitian. Furthermore, any Gramian matrix can be represented as $H = G^T G$ where $G = \left[v_1 \, v_2 \ldots v_n \right]$ is the matrix in which the i-th column is the vector v_i. Likewise, any matrix with the form $H = G^T G$ is Gramian. The properties of the submatrices of H are also important.

Def (1.3.1-2): The i,j-th submatrix of an m by n matrix A is the (m-1) by (n-1) matrix made by deleting the i-th row and j-th column of A. It is denoted $M_{i,j}$.

There is a special relationship between a Gramian matrix and its submatrices. Just as a Gramian matrix is made by taking the dot products of the set $v_1,\ldots,v_n$ with itself, the submatrix $M_{i,j}$ is made by taking the dot products of the set $v_1,\ldots v_{i-1},v_{i+1},\ldots,v_n$ with the set $v_1,\ldots v_{j-1},v_{j+1},\ldots,v_n$. This will become useful in later sections.

Theorem (1.3.1-3): Given a Gramian matrix $H = G^T G$ where $G = [v_1\, v_2 \ldots v_n]$, the submatrix $M_{i,j}$ can be written $M_{i,j} = G_i^T G_j$ where $G_k = [v_1 \ldots v_{k-1}\, v_{k+1} \ldots v_n]$.

Pf: Suppose that H is a Gramian matrix and let $M_{i,j}$ be an arbitrary submatrix of H. By definition 1.3.1-1, there is a set of vectors $v_1,\ldots,v_n$ such that $H_{ij} = v_i^T v_j$, $\forall_{1\le i,j\le n}$. Then writing H out we have:

$$H = \begin{bmatrix} \langle v_1, v_1\rangle & \langle v_1, v_j\rangle & \ldots & \langle v_1, v_n\rangle \\ \langle v_2, v_j\rangle & \langle v_2, v_2\rangle & \ldots & \langle v_2, v_n\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle v_n, v_1\rangle & \langle v_n, v_2\rangle & \ldots & \langle v_n, v_n\rangle \end{bmatrix}$$

By definition 1.3.1-2, the submatrix is then written with the i-th row and j-th column omitted.

$$M_{i,j} = \begin{bmatrix} \langle v_1, v_1 \rangle & \cdots & \langle v_1, v_{j-1} \rangle & \langle v_1, v_{j+1} \rangle & \cdots & \langle v_1, v_n \rangle \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \langle v_{i-1}, v_1 \rangle & \cdots & \langle v_{i-1}, v_{j-1} \rangle & \langle v_{i-1}, v_{j+1} \rangle & \cdots & \langle v_{i-1}, v_n \rangle \\ \langle v_{i+1}, v_1 \rangle & \cdots & \langle v_{i+1}, v_{j-1} \rangle & \langle v_{i+1}, v_{j+1} \rangle & \cdots & \langle v_{i+1}, v_n \rangle \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \langle v_n, v_1 \rangle & \cdots & \langle v_n, v_{j-1} \rangle & \langle v_n, v_{j+1} \rangle & \cdots & \langle v_n, v_n \rangle \end{bmatrix}$$

Now it can clearly be seen that the elements of $M_{i,j}$ are the dot products of the vectors

$v_1, \ldots v_{i-1}, v_{i+1}, \ldots, v_n$ in the rows and the vectors $v_1, \ldots v_{j-1}, v_{j+1}, \ldots, v_n$ in the columns.

Specifically, it can be written:

$$M_{i,j} = [v_1 \ldots v_{i-1}\, v_{i+1} \ \ldots v_n]^T [v_1 \ldots v_{j-1}\, v_{j+1} \ \ldots v_n] = G_i^T G_j$$

Q.E.D.

The most important case for investigating DOP is the case that $i = j$. In this special case, the submatrix is written as:

$$M_{i,i} = \begin{bmatrix} \langle v_1, v_1 \rangle & \cdots & \langle v_1, v_{i-1} \rangle & \langle v_1, v_{i+1} \rangle & \cdots & \langle v_1, v_n \rangle \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \langle v_{i-1}, v_1 \rangle & \cdots & \langle v_{i-1}, v_{i-1} \rangle & \langle v_{i-1}, v_{i+1} \rangle & \cdots & \langle v_{i-1}, v_n \rangle \\ \langle v_{i+1}, v_1 \rangle & \cdots & \langle v_{i+1}, v_{i-1} \rangle & \langle v_{i+1}, v_{i+1} \rangle & \cdots & \langle v_{i+1}, v_n \rangle \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \langle v_n, v_1 \rangle & \cdots & \langle v_n, v_{i-1} \rangle & \langle v_n, v_{i+1} \rangle & \cdots & \langle v_n, v_n \rangle \end{bmatrix}$$

or

$$M_{i,i} = [v_1 \ldots v_{i-1}\, v_{i+1} \ldots v_n]^T [v_1 \ldots v_{i-1}\, v_{i+1} \ldots v_n] = G_i^T G_i$$

This submatrix is clearly made by taking the dot products of the set $v_1, \ldots v_{i-1}, v_{i+1}, \ldots, v_n$ with itself. So, by definition 1.3.1-1, this submatrix is also a Gramian matrix. Thus, every submatrix M_{ii} of a Gramian is also a Gramian.

Corollary (1.3.1-4): The submatrix $M_{i,i}$ of a Gramian matrix $H = G^T G$ is itself Gramian and can be written in the form $M_{i,i} = G_i^T G_i$.

1.3.2 Parallelotopes and Volumes

Another special property of a Gramian matrix is that its determinant is related to a volume. Specifically, given a Gramian matrix $H = G^T G$, the determinant of H is the square of the volume of the parallelotope whose edges are the column vectors of the G matrix, and we write:

$$\det(H) = \det(G^T G) = Gram(v_1, \ldots, v_n) = Vol^2(v_1, \ldots, v_n) \qquad \text{Eq. 1.3.2-1}$$

The *Gram* notation indicates taking the determinant of the Gramian matrix formed using the given column vectors. The *Vol* notation indicates taking the volume of the parallelotope made up of the vectors. This and related results can be found in Gover and Krikorian (2010) or Jones

(2011). For the case in which $n = 3$ and $G = [v_1\ v_2\ v_3]$, the resulting parallelotope can be easily visualized.

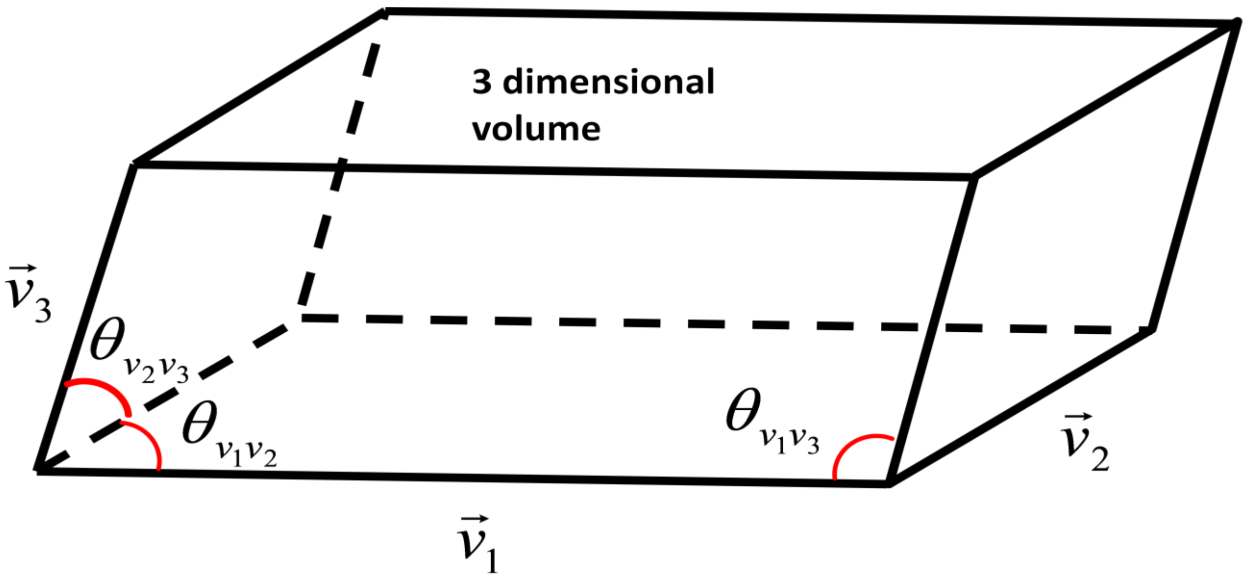


Figure 1.3.2-1

A three dimensional parallelotope is also known as a parallelpiped. A two dimensional parallelotope is simply a parallelogram. When H is a 2 by 2 Gramian matrix, the determinant is the square of the area of the resulting parallelogram. In general, the word “volume” will be used regardless of the fact that a parallelotope may have dimensions other than $n = 3$.

1.3.3 Properties of Parallelotopes

There is a special property of parallelotopes which can be established through the use of elementary column operations on the G matrix. There are three types of elementary column operations for matrices. Each can be performed by multiplying on the right by an elementary matrix. The elementary column operations and their corresponding matrices are listed here.

The determinant for each type of elementary matrix is also given.

1) Swapping the position of two columns in the matrix.

$$\begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 0 & & 1 & & \\ & & & \ddots & & & \\ & & 1 & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \quad \text{determinant = -1}$$

2) Adding a multiple of one column to another column.

$$\begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & \ddots & & & \\ & & c & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \quad \text{determinant = 1}$$

3) Multiplying any column by a non-zero scalar.

$$\begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & c & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \quad \text{determinant = c}$$

For any two square matrices A and B, $\det(AB) = \det(A)\det(B)$. So given an n by n Gramian matrix $H = G^T G$ and any n by n elementary column operation matrix E of the first or second type above:

$$\begin{aligned}\det((GE)^T(GE)) &= \det(E^T G^T GE) \\ &= \det(E^T)\det(G^T G)\det(E) \\ &= \det(E)^2 \det(G^T G) \\ &= \det(G^T G)\end{aligned}$$

This demonstrates that the columns of G can undergo elementary operations of the first two types without changing the determinant of $H = G^T G$. Since a parallelotope volume is simply the square root of a Gramian determinant, it inherits invariance under the same elementary column operations. So in general:

$$Vol(v_1, \ldots, v_n) = Vol(v_1, \ldots, v_{i-1}, v_i + av_j, v_{i+1}, \ldots, v_n) \qquad \text{Eq. 1.3.3-1}$$

for all $0 \le i, j \le n$ and any scalar a. Similarly,

$$Vol(v_1, \ldots, v_n) = Vol(v_1 + a_1 v_j, \ldots, v_{j-1} + a_{j-1} v_j, v_j, v_{j+1} + a_{j+1} v_j, \ldots, v_n + a_n v_j) \qquad \text{Eq. 1.3.3-2}$$

An interesting special case of equation 1.3.3-2 is when $a_i = \frac{< v_i, v_j >}{< v_j, v_j >}$ for each i. In this case

$$a_i v_j = \frac{< v_i, v_j >}{< v_j, v_j >} v_j = \Pr_{v_j}(v_i) \qquad \text{Eq. 1.3.3-3}$$

which is the projection of v_i onto v_j. Then

$$v_i - a_i v_j = v_i - \Pr_{v_j}(v_i) \qquad \text{Eq. 1.3.3-4}$$

which is orthogonal to v_i. Using this in equation 1.3.3-2 with $j = n$ yields:

$$Vol(v_1, \ldots, v_n) = Vol(v_1 - \Pr_{v_n}(v_1), v_2 - \Pr_{v_n}(v_2), \ldots, v_{n-1} - \Pr_{v_n}(v_{n-1}), v_n) \qquad \text{Eq. 1.3.3-5}$$

Since $v_i - \Pr_{v_n}(v_i)$ is orthogonal to v_n for all $1 \le i \le n-1$, this n dimensional volume can be reduced to an (n-1) dimensional volume.

$$Vol(v_1, \ldots, v_n) = \|v_n\| \cdot Vol(v_1 - \Pr_{v_n}(v_1), v_2 - \Pr_{v_n}(v_2), \ldots, v_{n-1} - \Pr_{v_n}(v_{n-1})) \qquad \text{Eq. 1.3.3-6}$$

This process could be repeated recursively to find the volume. It is equivalent to performing Gram-Schmidt orthogonalization on a determinant. That will not be the purpose here since such a process leads to formulas that become unwieldy even for n as small as 3.

Instead, a special case in which the G matrix contains a column of all ones is considered. This is an important case because such matrices arise very often in linear regression. The column of ones represents a constant or y-intercept term in the regression model. For this case the G matrix is written as $G = [v_1\ v_2\ \ldots\ v_{n-1}\ \mathbf{1}]$.[1] Using equation 1.3.3-6, the resulting volume is

$$Vol(v_1, \ldots, v_{n-1}, \vec{1}) = \|\mathbf{1}\| \cdot Vol(v_1 - \Pr_{\vec{1}}(v_1), v_2 - \Pr_{\vec{1}}(v_2), \ldots, v_{n-1} - \Pr_{\vec{1}}(v_{n-1})) \qquad \text{Eq. 1.3.3-7}$$

[1] Here the boldface 1 represents a column vector of ones, with dimension determined by context.

The norm of the ones vector is $\|\mathbf{1}\| = \sqrt{m}$ where m is the number of rows in G and the number of elements in each column vector. Also, subtracting a vector's projection onto **1** is equivalent to centering the vector since

$$\mathrm{Pr}_{\mathbf{1}}(v_i) = \frac{< v_i, \mathbf{1} >}{< \mathbf{1}, \mathbf{1} >} \cdot \mathbf{1} = \left(\frac{1}{m} \sum_{k=1}^{m} v_i(k) \right) \cdot \mathbf{1} = \bar{v} \cdot \mathbf{1}$$

Denoting the centered vector with a hat this yields

$$v_i - \mathrm{Pr}_{\bar{1}}(v_i) = v_i - \bar{v} \cdot \mathbf{1} = \hat{v}_i$$

Thus the volume for a G matrix including **1** is given by

$$Vol(v_1, \ldots, v_{n-1}, \mathbf{1}) = \sqrt{m} \cdot Vol(\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_{n-1}) \qquad \text{Eq. 1.3.3-8}$$

As a determinant equality this can be written

$$Gram(v_1, \ldots, v_{n-1}, \mathbf{1}) = m \cdot Gram(\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_{n-1}) \qquad \text{Eq. 1.3.3-9}$$

These equalities will prove to be useful in later sections.

1.3.4 DOP Using Parallelotopes

Recall that DOP terms are combinations of elements of the covariance matrix and that the covariance matrix Σ is the inverse of a Gramian matrix.

$$\Sigma = H^{-1} = (G'G)^{-1}$$

Since the covariance matrix is the inverse of a Gramian matrix, the inversion properties of Gramian matrices are of interest here. It is well known that the inverse of any non-singular square matrix is given by

$$\left(A^{-1}\right)_{i,j} = \frac{(-1)^{i+j}\left|M_{i,j}\right|}{|A|}$$ Eq. 1.3.4-1

where $M_{i,j}$ is the i,j-th submatrix of A. The determinant $\left|M_{i,j}\right|$ is called the i,j-th minor of A. This result can be found in basic linear algebra text books. For an easy proof see Nakos and Joyner (1998). The DOP terms are sums of elements on the main diagonal of the covariance matrix, so of special interest here are the terms

$$\Sigma_{i,i} = H_{i,i}^{-1} = \frac{(-1)^{2i}\left|M_{i,i}\right|}{|H|} = \frac{\left|M_{i,i}\right|}{|H|}$$ Eq. 1.3.4-2

From corollary 1.3.1-4, it is known that $M_{i,i}$ is a Gramian matrix by virtue of the fact that H is a Gramian matrix. Furthermore, equation 1.3.2-1 tells us that each of the determinants can be represented as a squared volume of a parallelotope, since they are both determinants of Gramian matrices. The result is:

$$\Sigma_{i,i} = \frac{\left|M_{i,i}\right|}{|H|} = \frac{Vol^2(v_1,\ldots,v_{i-1},v_{i+1},\ldots,v_n)}{Vol^2(v_1,\ldots,v_n)}$$ Eq. 1.3.4-3

Thus the diagonal elements of the covariance matrix are each a ratio of two volumes. It is important to note that since the numerator is the volume of the (n-1) dimensional parallelotope made up of $v_1,\ldots,v_{i-1},v_{i+1},\ldots,v_n$ and the denominator is the volume of the n dimensional parallelotope made up of $v_1, \ldots, v_n$, the first parallelotope is a "face" of the second. In a two dimensional case the larger volume is a parallelogram and the smaller volume is one edge (a one dimensional face) indicated below by blue.

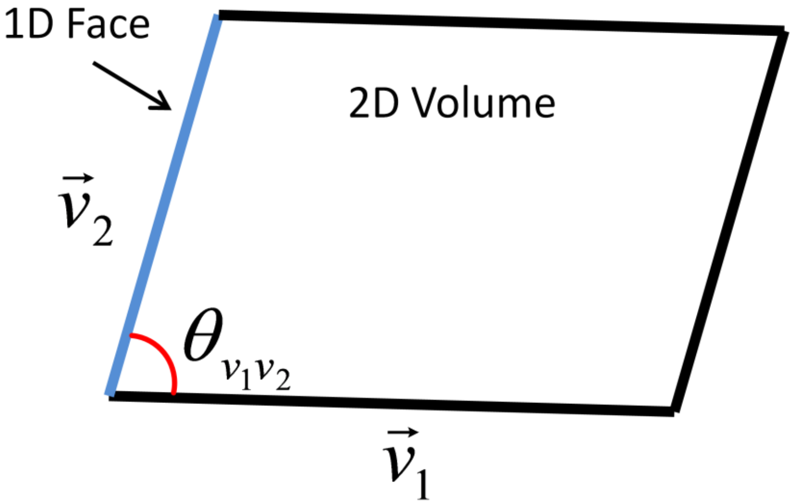


Figure 1.3.4-1

In the three dimensional case, the larger volume is a parallelpiped and the smaller volume is a parallelogram.

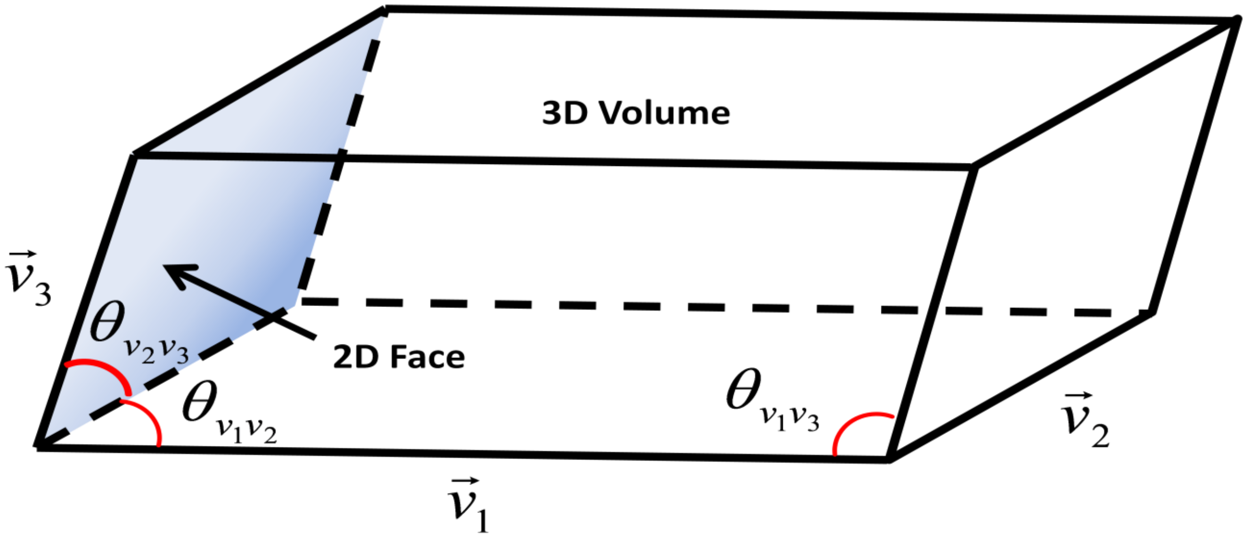


Figure 1.3.4-2

These concepts generalize in a natural way to higher dimensions. In general there is a hyper parallelotope of n dimensions with faces of (n-1) dimensions, however they become difficult to visualize for $n > 3$.

1.3.5 Additional Properties of Gramian Matrices and Parallelotopes

The parallelotope interpretation of the Gramian matrix presented in the previous section also extends to the inverse of the matrix. This section will lay out the relationship between a Gramian matrix, its inverse, and the related parallelotopes. First, it will be shown that the inverse of a Gramian matrix is also Gramian, which implies that there is a parallelotope associated with the inverse matrix as well. This will be a constructive proof and the construction will be important to understanding the inverse relationship.

Theorem 1.3.5-1: Given any non-singular Gramian matrix $H = G^T G$, the inverse H^{-1} is also Gramian.

pf: By definition 1.3.1-1, $H = G^T G$ for some m by n matrix $G = [v_1 \, v_2 \ldots v_n]$ where $v_1, \ldots, v_n \in \Re^m$ and $m \geq n$. Since H is non-singular it must be that $v_1, \ldots, v_n$ are linearly independent. Create a new set of linearly independent vectors $v_{n+1},\ldots,v_m \in \Re^m$, which are each orthogonal to the subspace of $\Re^m$ defined by $span(v_1,\ldots,v_n)$. Now create a new matrix $\Gamma = \left[v_1\, v_2 \ldots v_m\right] = [G \mid v_{n+1}\, v_{n+2} \ldots v_m]$. Because $< v_i, v_j >= 0$ whenever $v_i \in \{v_1,\ldots,v_n\}$ and $v_j \in \{v_{n+1},\ldots,v_m\}$, the associated Gramian matrix can be written:

$$\Gamma^T\Gamma = \begin{bmatrix} < v_1,v_1 > & \cdots & < v_1,v_n > & < v_1,v_{n+1} > & \cdots & < v_1,v_m > \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ < v_n,v_1 > & \cdots & < v_n,v_n > & < v_n,v_{n+1} > & \cdots & < v_n,v_m > \\ < v_{n+1},v_1 > & \cdots & < v_{n+1},v_n > & < v_{n+1},v_{n+1} > & \cdots & < v_{n+1},v_m > \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ < v_m,v_1 > & \cdots & < v_m,v_n > & < v_m,v_{n+1} > & \cdots & < v_m,v_m > \end{bmatrix} = \left[\begin{array}{c|c} G^TG & 0 \\ \hline 0 & \Gamma_2 \end{array}\right]$$

Now, a new set of vectors $\left[u_1\, u_2 \ldots u_m\right] \in \Re^m$ will be constructed using $\left[v_1\, v_2 \ldots v_m\right]$ in the following way. Consider any vector u_i. Since $v_1,\ldots,v_m$ are linearly independent, it must be possible to set the direction of u_i such that it is orthogonal to the subspace defined by $span(v_1,\ldots v_{i-1},v_{i+1},\ldots,v_m)$, but not orthogonal to v_i. Because u_i is not orthogonal to v_i, $< u_i, v_i > \neq 0$. So, set the length of u_i such that $< u_i, v_i >= 1$. Now form the matrix $\mathrm{B} = \left[u_1\, u_2 \ldots u_m\right]$. As before, the associated Gramian matrix can be written:

$$\mathrm{B}^T\mathrm{B} = \begin{bmatrix} < u_1,u_1 > & \cdots & < u_1,u_n > & < u_1,u_{n+1} > & \cdots & < u_1,u_m > \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ < u_n,u_1 > & \cdots & < u_n,u_n > & < u_n,u_{n+1} > & \cdots & < u_n,u_m > \\ < u_{n+1},u_1 > & \cdots & < u_{n+1},u_n > & < u_{n+1},u_{n+1} > & \cdots & < u_{n+1},u_m > \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ < u_m,u_1 > & \cdots & < u_m,u_n > & < u_m,u_{n+1} > & \cdots & < u_m,u_m > \end{bmatrix} = \left[\begin{array}{c|c} K^TK & \mathrm{B}_1^T \\ \hline \mathrm{B}_1 & B_2 \end{array}\right]$$

where $K = [u_1\, u_2 \ldots u_n]$.

Since u_i is orthogonal to $span(v_1, \ldots v_{i-1}, v_{i+1}, \ldots, v_m)$ it must be that $< u_i, v_j >= 0$ for all $i \neq j$. From $< u_i, v_i >= 1$and $< u_i, v_j >= 0$,

$$\Gamma^T \mathrm{B} = [v_1\, v_2 \ldots v_m]^T [u_1\, u_2 \ldots u_m] = I_{m,m}$$

so $\Gamma^T = \mathrm{B}^{-1}$ and

$$\Gamma \mathrm{B}^T = (\mathrm{B}\Gamma^T)^T = I_{m,m}$$

Thus

$$\mathrm{B}^T \mathrm{B} \Gamma^T \Gamma = \mathrm{B}^T \mathrm{B} \mathrm{B}^{-1} \Gamma = \mathrm{B}^T \Gamma = I_{m,m}$$

Using the partitioned forms of $\mathrm{B}^T\mathrm{B}$, $\Gamma^T\Gamma$, and $I_{m,m}$ yields:

$$\left[\begin{array}{c|c} I_{n,n} & 0 \\ \hline 0 & I_{m-n,m-n} \end{array}\right] = \left[\begin{array}{c|c} K^T K & \mathrm{B}_1^T \\ \hline \mathrm{B}_1 & \mathrm{B}_2 \end{array}\right] \left[\begin{array}{c|c} G^T G & 0 \\ \hline 0 & \Gamma_2 \end{array}\right]$$

Multiplying partitioned matrices, the equation becomes:

$$\left[\begin{array}{c|c} I_{n,n} & 0 \\ \hline 0 & I_{m-n,m-n} \end{array}\right] = \left[\begin{array}{c|c} K^T K G^T G + \mathrm{B}_1^T 0 & K^T K 0 + \mathrm{B}_1^T \Gamma_2 \\ \hline \mathrm{B}_1 G^T G + \mathrm{B}_2 0 & \mathrm{B}_1 0 + \mathrm{B}_2 \Gamma_2 \end{array}\right] = \left[\begin{array}{c|c} K^T K G^T G & \mathrm{B}_1^T \Gamma_2 \\ \hline \mathrm{B}_1 G^T G & \mathrm{B}_2 \Gamma_2 \end{array}\right]$$

From the top left partition, it can be seen that

$$I_{n,n} = K^T K G^T G = LH$$

where $L = K^T K$. Thus a matrix $L = K^T K$ has been constructed which is Gramian by definition and which is the inverse of H. The inverse is necessarily unique for any non-singular matrix. Thus the inverse of a Gramian matrix is itself Gramian.

Q.E.D.

In practice, the direction of each u_i could be found by taking the cross product $v_1 \times \cdots \times v_{i-1} \times v_{i+1} \times \cdots \times v_m$, which is a vector that is orthogonal to the subspace defined by $span(v_1, \ldots v_{i-1}, v_{i+1}, \ldots, v_m)$. It should also be noted that when $m = n$ in the above proof the set $v_{n+1}, \ldots, v_m$ is empty, $\Gamma = G$, and $\mathrm{B} = K$.

The theorem can also be proven using the fact that a nonsingular matrix is Gramian if and only if it is both positive definite and symmetric. It can be shown that the inverse is positive definite and symmetric also. Such a proof, based on the spectral representation of the matrix, may be analytically more simple, however it will not provide a construction for the matrix K with special geometric properties as the above proof does.

From the construction of K in the above proof, there is an interpretation of the matrix inverse as a set of parallelotopes. The G matrix is made up of a set of vectors which form a parallelotope.

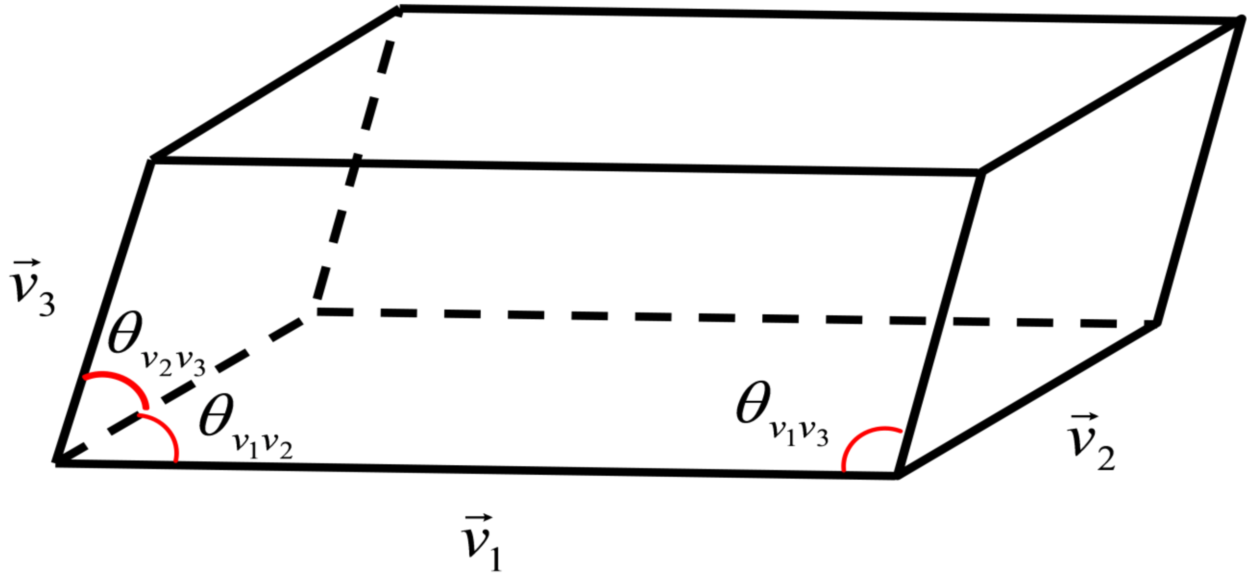


Figure 1.3.5-1

Since each u_i is orthogonal to $span(v_1,\ldots v_{i-1},v_{i+1},\ldots,v_n)$, it is orthogonal to the (n-1) dimensional parallelotope face made up of $v_1,\ldots v_{i-1},v_{i+1},\ldots,v_n$. When $m=n$ this direction is unique for fixed $v_1,\ldots,v_n$. When $m\geq n$ the vector u_i is still orthogonal to the face, but the direction of u_i is not unique because the face does not span an (m-1) dimensional subspace of $\Re^m$. If the process in the above proof is followed, the direction of u_i will be determined by the choice of $v_{n+1},\ldots,v_m$.

By virtue of the fact that $<u_i,u_i>$ is the i-th diagonal term of the matrix H^{-1}, from equation 1.3.4-2,

$$\|u_i\| = \sqrt{<u_i,u_i>} = \sqrt{\frac{|M_{i,i}|}{|H|}} \qquad \text{Eq. 1.3.5-1}$$

This is the volume of the face divided by the volume of the entire parallelotope formed by G. The relationship between the constructed vectors $u_1,\ldots,u_n$ and the vectors $v_1,\ldots,v_n$ is pictured below.

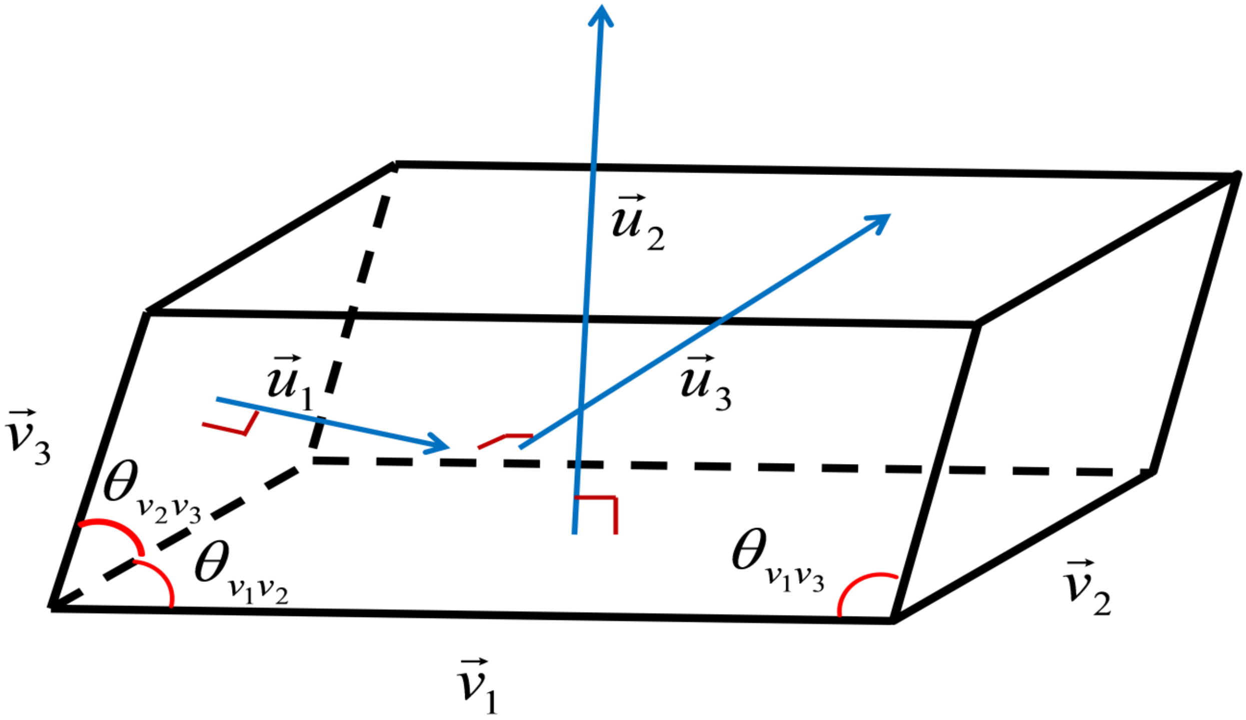


Figure 1.5.3-2

Although the inverse matrix L is unique, the matrix K is not. In fact, the matrix G which forms the unique matrix $H = G^T G$ is not itself a unique matrix. Notably, the matrix H is invariant under rotations of G.

Def 1.3.5-2: A matrix R is a rotation[2] matrix if $R^T = R^{-1}$ and $\det(R) = 1$.

Such matrices cause a rotation of the vectors in a matrix A when A is multiplied by R.

Theorem 1.3.5-2: A Gramian matrix $H = G^T G$ is invariant under rotation of the matrix G.

pf: Substitute G with any rotation RG and

$$(RG)^T (RG) = G^T R^T RG = G^T R^{-1} RG = G^T G = H$$

Q.E.D.

Seeing G as a set of vectors, one might view the action of taking $H = G^T G$ as throwing out the rotational information of the vectors and keeping the moment information. So in actuality, the vectors $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ need not be orthogonal in the way they are shown in Figure 1.3.5-3 since either set may be rotated without changing H or L. However, the interior angles between the vectors $u_1, \ldots, u_n$ still represent the principal angles between the faces of the parallelotope made up by $v_1, \ldots, v_n$. It should be noted that this goes both ways. Since H is also the inverse of L, the vectors $u_1, \ldots, u_n$ can be viewed as a paralleltope and the vectors $v_1, \ldots, v_n$ as the orthogonal vectors to the faces. So there are two related

[2] Also called an "orthogonal matrix".

parallelotopes. The interior angles of one are the principal angles of the other and vice versa. The edge lengths of one are related to the face volumes of the other by equation 1.3.5-1.

Each piece of information on the original parallelotope corresponds to a term in the H matrix. The edge lengths correspond to the diagonal terms. For example $H_{1,1} = \|v_1\|^2$. The interior angles are related to the off diagonal terms of H following the definition of dot product. So the figures above show a relationship between the original matrix H and the covariance matrix $\Sigma = H^{-1}$. H is represented by the edge lengths and interior angles of the parallelotope. The covariance matrix Σ is represented by face-volumes and their principal angles. The determinant $|H| = |\Sigma|^{-1}$ is represented by the square of the total volume of the parallelotope.

1.3.6 Using Centered Vectors

When the vectors associated with a parallelotpe are centered, that is their components sum to zero, the parallelotope has special relationships with the statistical properties of the vectors. This can be helpful in understanding some dilution of precision problems and it arises in any case in which the associated linear regression model includes an intercept. Two lemmas with simple proofs are given here. The first describes the relationship between a centered vector's norm and the vector's sample standard deviation. The second describes the relationship between the interior angle of two centered vector's and the sample correlation of the vectors. Both of these lemmas are basic but useful.

Lemma (1.3.5-1): $\|\hat{x}\| = \sqrt{m} \cdot \sigma_x$

pf: $\|\hat{x}\| = \sqrt{< \hat{x}, \hat{x} >}$

$$= \sqrt{< x - \bar{x} \cdot \mathbf{1}, x - \bar{x} \cdot \mathbf{1} >}$$

$$= \sqrt{< x, x > -2 < x, \bar{x} \cdot \mathbf{1} > + < \bar{x} \cdot \mathbf{1}, \bar{x} \cdot \mathbf{1} >} \qquad \text{by bilinearity}$$

$$= \sqrt{< x, x > -2\bar{x} < x, \mathbf{1} > + \bar{x}^2 < \mathbf{1}, \mathbf{1} >}$$

$$= \sqrt{\sum_{i=1}^{m} x_i^2 - 2\bar{x}\sum_{i=1}^{m} x_i + \bar{x}^2 \sum_{i=1}^{m} 1}$$

$$= \sqrt{\sum_{i=1}^{m} x_i^2 - 2m \cdot \bar{x}^2 + m \cdot \bar{x}^2}$$

$$= \sqrt{m} \cdot \sqrt{\frac{1}{m}\left(\sum_{i=1}^{m} x_i^2\right) - \bar{x}^2}$$

$$= \sqrt{m} \cdot \sqrt{\frac{1}{m}\left(\sum_{i=1}^{m} x_i^2\right) - \left(\frac{1}{m}\sum_{i=1}^{m} x_i\right)^2}$$

$$= \sqrt{m} \cdot \sigma_{\vec{x}}$$

Q.E.D.

So, the norm of the centered vector is $\sqrt{m}$ times the sample standard deviation of the non-centered vector (which is also the standard deviation of the centered vector). This describes the edge lengths of a parallelotope made up of centered vectors.

Lemma (1.3.6-2): $\cos(\theta(\hat{x},\hat{y})) = \rho_{x,y}$

pf: $\cos(\theta(\hat{x},\hat{y})) = \dfrac{<\hat{x},\hat{y}>}{\|\hat{x}\|\cdot\|\hat{y}\|}$

$$= \frac{< x-\bar{x}\cdot\mathbf{1}, y-\bar{y}\cdot\mathbf{1} >}{\|\hat{x}\|\cdot\|\hat{y}\|}$$

$$= \frac{< x,y > - < \bar{x}\cdot\mathbf{1}, y > - < x, \bar{y}\cdot\mathbf{1} > + < \bar{x}\cdot\mathbf{1}, \bar{y}\cdot\mathbf{1} >}{\|\hat{x}\|\cdot\|\hat{y}\|}$$

$$= \frac{< x,y > -\bar{x} < \mathbf{1}, y > -\bar{y} < x,\mathbf{1} > +\bar{x}\cdot\bar{y} < \mathbf{1},\mathbf{1} >}{\|\hat{x}\|\cdot\|\hat{y}\|}$$

$$= \frac{< x,y > -m\cdot\bar{x}\cdot\bar{y} - m\cdot\bar{y}\cdot\bar{x} + m\cdot\bar{x}\cdot\bar{y}}{\|\hat{x}\|\cdot\|\hat{y}\|}$$

$$= \frac{< x,y > -m\cdot\bar{x}\cdot\bar{y}}{\|\hat{x}\|\cdot\|\hat{y}\|}$$

$$= \frac{\sum_{i=1}^{m} x_i y_i - \frac{1}{m}\cdot\sum_{i=1}^{m} x_i\cdot\sum_{i=1}^{m} y_i}{\|\hat{x}\|\cdot\|\hat{y}\|}$$

$$= \frac{m \cdot \left(\frac{1}{m} \sum_{i=1}^{m} x_i y_i - \left(\frac{1}{m} \sum_{i=1}^{m} x_i \right) \cdot \left(\frac{1}{m} \sum_{i=1}^{m} y_i \right) \right)}{\|\hat{x}\| \cdot \|\hat{y}\|}$$

$$= \frac{m \cdot \operatorname{cov}(x, y)}{\|\hat{x}\| \cdot \|\hat{y}\|}$$

$$= \frac{m \cdot \operatorname{cov}(x, y)}{\sqrt{m} \cdot \sigma_x \cdot \sqrt{m} \cdot \sigma_y}$$ from the previous result

$$= \rho_{x,y}$$

Q.E.D.

So the cosine of the interior angle of the centered vectors is the correlation of the corresponding non-central vectors. This describes the interior angles of a parallelotope made up of centered vectors.

1.3.7 Volumes of Paralellotopes

As seen in equation 1.3.2-1, the general volume of an N dimensional parallelotope can be represented as the determinant of a the Gramian matrix $H = G^T G$ where the set of column vectors $G = [v_1 \, v_2 \ldots v_n]$ are the vectors which make up the edges of the parallelotope. Another way to represent the general volume for parallelotopes is with exterior products.

$$Vol(v_1,\ldots,v_n)=\|v_1\wedge\ldots\wedge v_n\|$$

However, the exterior products will not be used here. As shown by Gram-Schmidt orthogonalization in section 1.3.3 and particularly in equation 1.3.3-6, these are equivalent to using a Gram-Schmidt process or repeated sine rule. That is to say that starting with the volume as the length of the first vector, each successive vector scales the volume by the length of the vector times the sine of the angle between that vector and the current volume.

For lower dimensional cases, the specific equations for volume are already known. For the two dimensional case, which is simply a parallelogram, the 2D-volume is the area and the formula is simply:

$$\begin{aligned}Area(\vec{x}_1,\vec{x}_2)&=\|\vec{x}_1\|\cdot\|\vec{x}_2\|\cdot\sqrt{1-\cos^2(\theta(\vec{x}_1,\vec{x}_2))}\\&=\|\vec{x}_1\|\cdot\|\vec{x}_2\|\cdot\sqrt{\sin^2(\theta(\vec{x}_1,\vec{x}_2))}\end{aligned}\qquad\text{Eq. 1.3.7-1}$$

where $\theta(\vec{x},\vec{y})$ is the interior angle between the vectors $\vec{x}$ and $\vec{y}$. In the three dimensional case, the parallelpiped has 3D-volume given by:

$$\begin{aligned}Vol(\vec{x}_1,\vec{x}_2,\vec{x}_3)=\|\vec{x}_1\|\cdot\|\vec{x}_2\|\cdot\|\vec{x}_3\|\cdot[1+2\cos(\theta(\vec{x}_1,\vec{x}_2))\cos(\theta(\vec{x}_2,\vec{x}_3))\cos(\theta(\vec{x}_1,\vec{x}_3))-\\ \cos^2(\theta(\vec{x}_1,\vec{x}_2))-\cos^2(\theta(\vec{x}_2,\vec{x}_3))-\cos^2(\theta(\vec{x}_1,\vec{x}_3))]^{1/2}\end{aligned}\qquad\text{Eq. 1.3.7-2}$$

(Gover & Krikorian, 2010). Both of these formulas are also found in Jones (2011).

In the case that the parallelotope is made up of centered vectors, the formulae have centered forms. From lemmas 1.3.6-1, 1.3.6-2 and equation 1.3.7-1 the area for a parallelogram made of centered vectors is:

$$Area(\hat{x}_1, \hat{x}_2) = m \cdot \sigma_{\vec{x}_1} \cdot \sigma_{\vec{x}_2} \cdot \sqrt{1 - \rho^2{}_{\vec{x}_1, \vec{x}_2}} \qquad \text{Eq. 1.3.7-3}$$

From lemmas 1.3.6-1, 1.3.6-2 and equation 1.3.7-2, the volume for a parallelpiped made up of centered vectors is:

$$\begin{aligned} Vol(\hat{x}_1, \hat{x}_2, \hat{x}_3) = m^{3/2} \cdot \sigma_{\hat{x}_1} \cdot \sigma_{\hat{x}_2} \cdot \sigma_{\hat{x}_3} \cdot [1 + 2\rho_{\vec{x}_1, \vec{x}_2} \rho_{\vec{x}_2, \vec{x}_3} \rho_{\vec{x}_1, \vec{x}_3} - \\ \rho^2_{\vec{x}_1, \vec{x}_2} - \rho^2_{\vec{x}_2, \vec{x}_3} - \rho^2_{\vec{x}_1, \vec{x}_3}]^{1/2} \end{aligned} \qquad \text{Eq. 1.3.7-4}$$

These centered volume equations will become useful in section 1.4.

1.3.8 A-Optimality and D-Optimality

DOP optimization using parallelotopes is highly related to the concepts of A-optimality and D-optimality for matrices. These are well known optimality ideas for regression models that are based on the information matrix G and covariance matrix $G^T G$. Descriptions for each and discussions about their existence can be found in Rao (1965), Chan and Wong (1980), Chan (1982), and Hooks et al. (2009).

Def (1.3.8-1): An A-optimal solution for a linear regression design problem with design matrix G is the choice of G that minimizes the trace of $G^T G$ which is denoted by $tr(G^T G)$.

Since the trace is the sum of the diagonal elements of a matrix, $tr((G^T G)^{-1})$ is proportional to the sum of the variances of each least squares estimate in the linear regression problem. For GPS, it will be seen that an A-optimal solution is exactly the optimal solution for GDOP. Other DOPs are similar, but are each a sum over only a subset of the diagonal elements of $G^T G$. The exact solutions for these DOPs will be called "A-optimal type" solutions.

Def (1.3.8-2): A D-optimal solution for a linear regression design problem with design matrix G is the choice of G that minimizes $\det((G^T G)^{-1})$.[3]

In general, the inverse of the determinant of a matrix is the determinant of the inverse. So, this is equivalently the maximization of $\det(G^T G)$. From equation 1.3.2-1, this is the same as the maximization of $Vol(v_1,\ldots,v_n)$. So the D-optimal solution(s) occur when the parallelotope associated with the control matrix G achieves a maximum volume. From equation 1.3.4-3, it can be seen that a D-optimal solution for G maximizes the denominator of any DOP term ignoring the numerator, whereas an A-optimal type solution minimizes the actual DOP. While A-optimal type solutions will be the focus of the following sections, a D-optimal solution will be discussed in section 1.4.1.

[3] In multivariate analysis, this is called the generalized variance, and minimizing it is equivalent to minimizing the geometric mean of the eigenvalues of the inverse of G'G, which are themselves proportional to the variances of the principal components of the G matrix.

1.4 DOP Optimization for GPS

This section will use the developments of section 1.3 to demonstrate how to find optimal geometries for various DOPs in the GPS system. These findings include conceptually based interpretations along with exact matching formulas and discussions about the equivalence of the two.

1.4.1 Known DOP Results for GPS

There are some previously known results for optimal DOP in GPS. A solution is presented for the optimal configuration for GDOP given exactly 4 satellites in Parkinson and Spilker (1996). The argument is that GDOP is made up of a numerator and denominator term, as we will see in the following sections. Since the denominator has higher order terms, this portion is maximized and the numerator is ignored in an effort to minimize the ratio. Because of this, the solution will be an estimate. Since the denominator is the determinant of $G^T G$, it can now be seen that this solution is exactly a D-optimal solution as per definition 1.3.8-2. It is determined that the determinant, and thus the denominator, is maximized when the 4 satellites form a tetrahedron with one satellite at 90 degree elevation and the others equally spaced at 0 degree elevation. So this is the D-optimal satellite geometry for GDOP with exactly 4 satellites. The optimal tetrahedron geometry is also discussed in Marquis (1993) and Langley (1999).

Although the following sections of 1.4 will focus on A-optimal type solutions, from the results to follow one might speculate that this is a very good estimate and that for 4 satellites a tetrahedron geometry is desirable. However, one must be careful with D-optimal solutions for GPS. All DOPs share the same denominator as GDOP. Also, they are all bounded above by GDOP. So, the exact same argument could be used to find the same D-optimal solution for any DOP given 4 satellites, which would not be the correct approach. In particular, this would say that VDOP and HDOP are optimized for the same satellite configuration, which makes no sense. However, the A-optimal type solutions in general for various DOPs, most notably VDOP and HDOP, will be very different from each other. So it is important to find the exact A-optimal type solutions. Furthermore, these solutions will work for any number of satellites m when $m \geq 4$.

Parkinson & Spilker (1996) also contains computer based results for GDOP with more than 4 satellites and a discussion of DOP for the two dimensional case. For GPS purposes here, only three dimensional cases will be considered.

1.4.2 Parallelotopes and Vertical DOP (VDOP)

Using the DOP matrix $G = [e\ n\ u\ \mathbf{1}]$ for GPS and equation 1.3.4-3, VDOP can be viewed as the ratio of the two parallelotope volumes:

$$VDOP = \sqrt{\Sigma_{u,u}} = \sqrt{\frac{|M_{u,u}|}{|H|}} = \frac{Vol(e,n,\mathbf{1})}{Vol(e,n,u,\mathbf{1})} \qquad \text{Eq. 1.4.2-1}$$

It may not be necessary, but in this case it will be helpful to use equation 1.3.3-8 to centralize the coordinate vectors and to lower the number of dimensions in the parallelotopes each by one. Then,

$$VDOP = \frac{Vol(\hat{e},\hat{n})}{Vol(\hat{e},\hat{n},\hat{u})} \qquad \text{Eq. 1.4.2-2}$$

As discussed in section 1.3.4, the parallelotope represented in the top numerator is a face of the parallelotope represented by the denominator. So there is a visualization similar to Figure 1.3.4-2 specifically for VDOP.

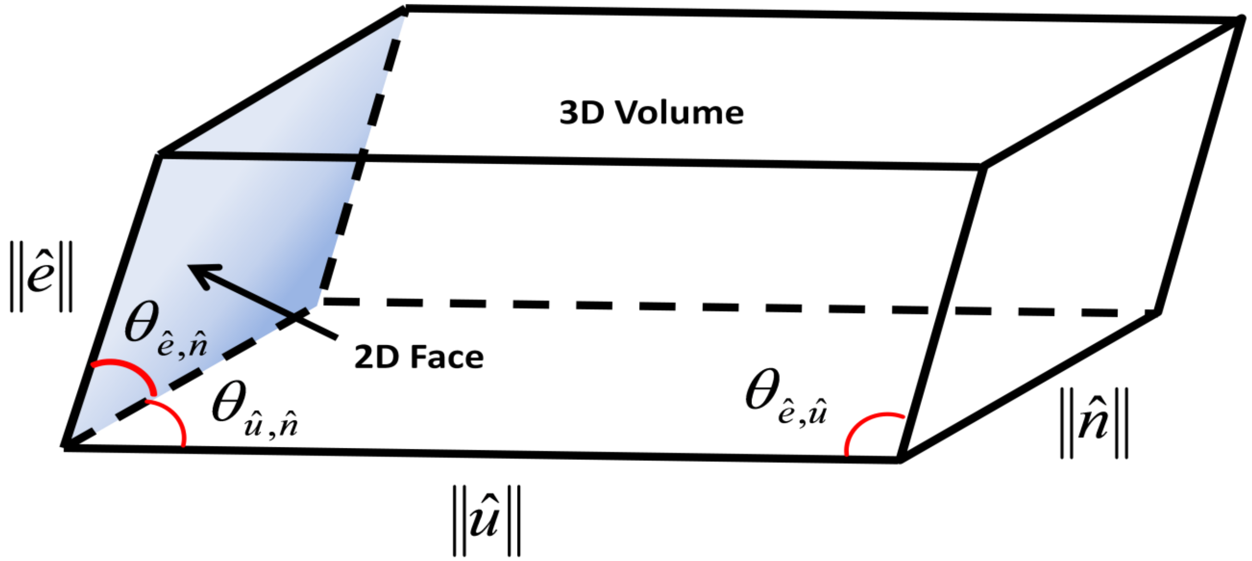

3D Volume
$\|\hat{e}\|$
$\theta_{\hat{e},\hat{n}}$
$\theta_{\hat{u},\hat{n}}$
2D Face
$\theta_{\hat{e},\hat{u}}$
$\|\hat{n}\|$
$\|\hat{u}\|$

Figure 1.4.2-1

Had Equation 1.3.3-8 not been used, there would instead be a 4-dimensional hypertope with non-central vector edges. From lemmas 1.3.6-1 and 1.3.6-2, another representation of the above figure is:

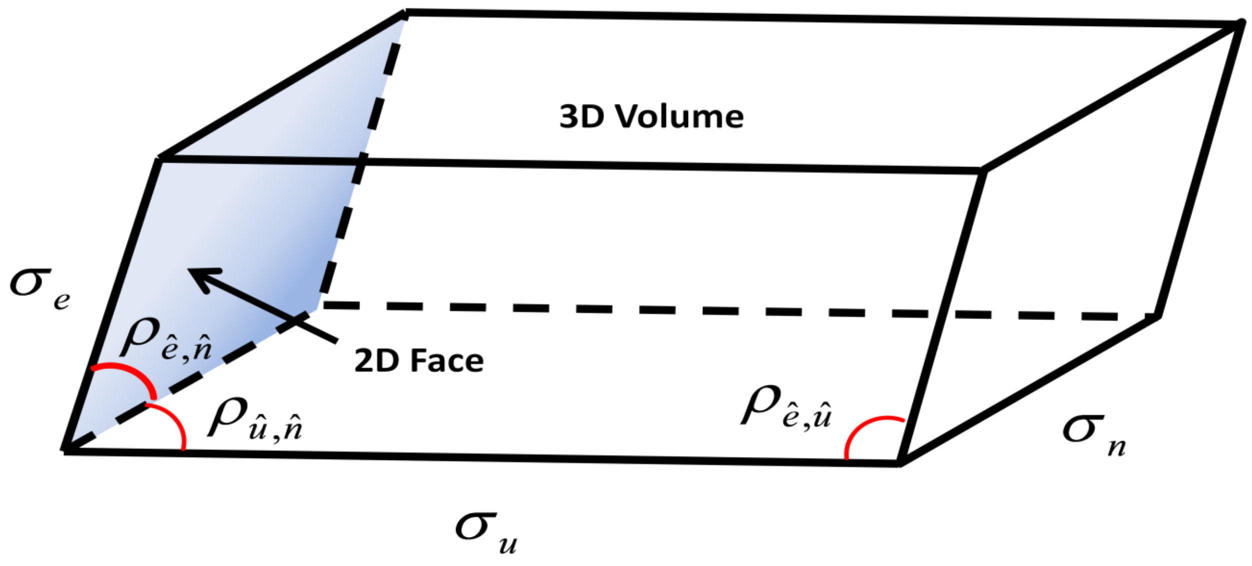


Figure 1.4.2-2

According to lemma 1.3.6-1, each edge should also have a $\sqrt{m}$, however these will turn out to be irrelevant in minimizing DOP for a fixed number of satellites. Here, they are left out for simplicity, but their exact effect on DOP will be seen in the following sections.

An exact minimization for VDOP can be determined by examination of the above figure. Consider the edge σ_e. If this edge is expanded or contracted, both the 2D face and the 3D volume are increased or decreased by the same proportion. As a result, the ratio of the two volumes or the VDOP is unchanged by this factor. So, σ_e has no effect on VDOP. Similarly, σ_n also has no effect. However, σ_u has an important effect on VDOP. Increasing σ_u increases the 3D volume but has no effect on the 2D face. So to minimize the ratio it is desirable to have a large σ_u. As σ_u decreases to zero, VDOP will approach infinity. Because the 3D volume increases linearly as σ_u increases, the relationship between σ_u and VDOP is inverse linear.

This is a complete description of the standard deviations or edge lengths that are required for optimization.

It remains to find the optimal configuration for the correlations. Clearly, the ratio of the two volumes is smallest when the σ_u edge is orthogonal to the 2D face. The denominator is proportional to the cosine of the angle between the σ_u edge and the 2D face. However the numerator is unaffected by this angle. To achieve orthogonality, we should have $\rho_{\hat{n},\hat{u}} = \rho_{\hat{e},\hat{u}} = 0$. When this occurs, the term $\rho_{\hat{e},\hat{n}}$ has no effect on the ratio because each volume is proportional to the cosine of the interior angle for the 2D face. In general, $\rho_{\hat{e},\hat{n}}$ does have an effect on the ratio, but not when the other correlations are zero.

So it can be seen that the ratio of the volumes, and hence VDOP, is minimized when σ_u is maximized and when $\rho_{\hat{n},\hat{u}} = \rho_{\hat{e},\hat{u}} = 0$. However, there is not necessarily a way to know that both of these can occur at the same time for a given problem, unless an example solution is found which satisfies both. In the case of VDOP it is easy to find such a configuration for the satellites. One possible configuration for $m = 8$ is:

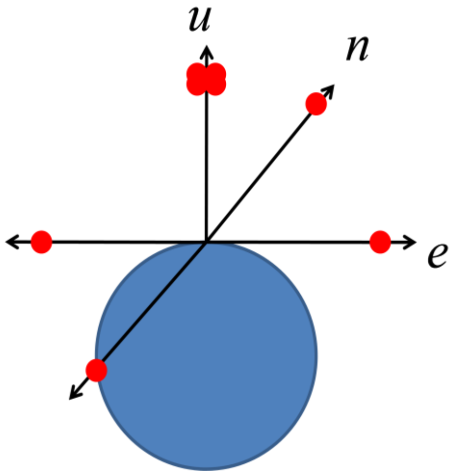


Figure 1.4.2-3

This figure shows 4 satellites collocated directly overhead, and 4 equally spaced at 0 degrees elevation. In this case, the variance in the u direction is maximized because half of the u coordinates are at the upper bound and half are at the lower bound. The correlations $\rho_{\hat{e},\hat{u}}$ and $\rho_{\hat{n},\hat{u}}$ are both zero by symmetry. So Figure 1.4.2-3 represents an optimal configuration of satellites for VDOP. From the discussion of this section and from Figure 1.4.2-2, the effects on VDOP of changing σ_u , $\rho_{\hat{e},\hat{u}}$, or $\rho_{\hat{n},\hat{u}}$ are known.

1.4.3 The VDOP Formula

This analysis leads directly to the exact formula for VDOP in terms of standard deviations and correlations. It can be useful to look at the algebraic formula and it can be seen that the results correspond to those found by examining the parallelotope above. From simply applying equations 1.3.7-3 and 1.3.7-4 to equation 1.4.2-2, the ratio for VDOP becomes:

$$VDOP = \frac{\sigma_{\vec{e}} \cdot \sigma_{\vec{n}} \cdot \sqrt{1 - p_{\vec{e},\vec{n}}^2}}{\sqrt{m} \cdot \sigma_{\vec{e}} \cdot \sigma_{\vec{n}} \cdot \sigma_{\vec{u}} \cdot \sqrt{1 + 2 p_{\vec{e},\vec{n}} p_{\vec{n},\vec{u}} p_{\vec{e},\vec{u}} - p_{\vec{e},\vec{n}}^2 - p_{\vec{n},\vec{u}}^2 - p_{\vec{e},\vec{u}}^2}} \qquad \text{Eq. 1.4.3-1}$$

Immediately, we see as in the visualization, that the standard deviations of the east and north directions cancel. It is also clear that VDOP indeed has an inverse linear relationship to the standard deviation in the up direction. The result is:

$$VDOP = \frac{\sqrt{1-p_{\vec{e},\vec{n}}^2}}{\sqrt{m}\cdot\sigma_{\vec{u}}\cdot\sqrt{1+2p_{\vec{e},\vec{n}}p_{\vec{n},\vec{u}}p_{\vec{e},\vec{u}}-p_{\vec{e},\vec{n}}^2-p_{\vec{n},\vec{u}}^2-p_{\vec{e},\vec{u}}^2}} \qquad \text{Eq. 1.4.3-2}$$

This is the VDOP formula. It is of interest to consider that this term should be real by virtue of the fact that it is a standard deviation. It is easy to see that the numerator is indeed real. A proof that the term under the radical in the denominator is non-negative can be found in Jones (2011). No further proof that the denominator is real is offered here. Another important note is that the formula involves only the standard deviation and correlation terms in addition to the vector size m and that these are exactly the pieces of information that existed in the parallelotope analysis. This implies that all the relevant information is indeed on the parallelotope. This will always hold for parallelotope analysis because the parallelotope contains every piece of information that is in the Gramian matrix, which is the only input to the DOP problem.

In a case such as this, the minimum can be identified by treatment of the DOP formula. In higher dimensional cases, the formulae may quickly become too large to be practical making the parallelotope visualization even more useful. Minimization via formula will be performed here for VDOP. Although it is possible to approach this using the typical method of setting the partial derivatives of the function to zero, those being derivates with respect to ($p_{\vec{e},\vec{n}}$, $p_{\vec{n},\vec{u}}$, $p_{\vec{e},\vec{u}}$, and $\sigma_{\vec{u}}$), a different method will be used here which does not involve calculus. If using the derivative method, it is perhaps easiest to minimize $VDOP^2$ instead. As an alternative, we simply divide the numerator and denominator of the formula both by $\sqrt{1-p_{\vec{e},\vec{n}}^2}$. This yields:

$$VDOP = \frac{1}{\sqrt{m} \cdot \sigma_{\vec{u}} \cdot \sqrt{\frac{2 p_{\vec{e},\vec{n}} p_{\vec{n},\vec{u}} p_{\vec{e},\vec{u}} - p_{\vec{n},\vec{u}}^2 - p_{\vec{e},\vec{u}}^2}{1 - p_{\vec{e},\vec{n}}^2} + 1}}$$

The fractional term under the radical can be shown to be non-positive. In fact, it is bounded inside [-1,0]. I.e.,

$$-1 \leq \frac{2 p_{\vec{e},\vec{n}} p_{\vec{n},\vec{u}} p_{\vec{e},\vec{u}} - p_{\vec{n},\vec{u}}^2 - p_{\vec{e},\vec{u}}^2}{1 - p_{\vec{e},\vec{n}}^2} \leq 0$$

A proof of the right side is included in the appendix. The left side follows from the proof in Jones (2011) mentioned above. This fact tells us that VDOP is bounded below by the function:

$$VDOP \geq \frac{1}{\sqrt{m} \cdot \sigma_{\vec{u}}}$$

It is easy to see that this bound can be achieved whenever the correlation terms are all zero. In fact, it is achieved whenever $p_{\vec{n},\vec{u}} = p_{\vec{e},\vec{u}} = 0$. When this happens the $p_{\vec{e},\vec{n}}$ term is inconsequential. This corresponds to exactly what is seen in the parallelotope analysis, that when the two interior angels matching $p_{\vec{n},\vec{u}}$ and $p_{\vec{e},\vec{u}}$ are right angles, the remaining angle $p_{\vec{e},\vec{n}}$ has no bearing on the ratio of the volumes.

As before, it is apparent that the standard deviation can be maximized while the correlations are simultaneously zeroed. So the same minimal solution is achieved as in the parallelotope analysis. As we would expect, the formulaic and visual analysis yield the same results and have the same characteristics. With the visualization, however, it is easier to see the real behavior of the function and it is easier to consider higher dimensional cases. An example of this will be seen in the TDOP section below.

1.4.4 NDOP and EDOP Formulas and Minimization

The exact same process used for VDOP yields an identical parallelotope analysis for NDOP and EDOP along with DOP equations of the same form.

$$NDOP = \frac{\sqrt{1 - p_{\vec{e},\vec{u}}^2}}{\sqrt{m} \cdot \sigma_{\vec{n}} \cdot \sqrt{1 + 2 p_{\vec{e},\vec{n}} p_{\vec{n},\vec{u}} p_{\vec{e},\vec{u}} - p_{\vec{e},\vec{n}}^2 - p_{\vec{n},\vec{u}}^2 - p_{\vec{e},\vec{u}}^2}} \qquad \text{Eq. 1.4.4-1}$$

$$EDOP = \frac{\sqrt{1 - p_{\vec{n},\vec{u}}^2}}{\sqrt{m} \cdot \sigma_{\vec{e}} \cdot \sqrt{1 + 2 p_{\vec{e},\vec{n}} p_{\vec{n},\vec{u}} p_{\vec{e},\vec{u}} - p_{\vec{e},\vec{n}}^2 - p_{\vec{n},\vec{u}}^2 - p_{\vec{e},\vec{u}}^2}} \qquad \text{Eq. 1.4.4-2}$$

However, there is a difference due to the realities of GPS. For a user on the globe, a satellite may be in the North or South direction which represent positive North and negative North coordinates. Similarly it can be in the positive East or negative East direction. So the normalized North and East coordinates each fall in [-1,1]. This is different from the up coordinates which fall in [0,1] due to the fact that any satellite in the negative up direction is below the horizon and is thus out of view from a user on the ground. It should be noted that it is possible for satellites to be in view and below some users, such as airplanes using GPS guidance. In this case the up coordinates are not bounded below by zero.

Since the parallelotopes and formulas are the same as in section 1.4.2 and 1.4.3, the minimization is still achieved by maximizing the variance and zeroing the correlations in the direction of interest. As in the VDOP case, these can be done simultaneously. For $m = 8$, the configurations would be:

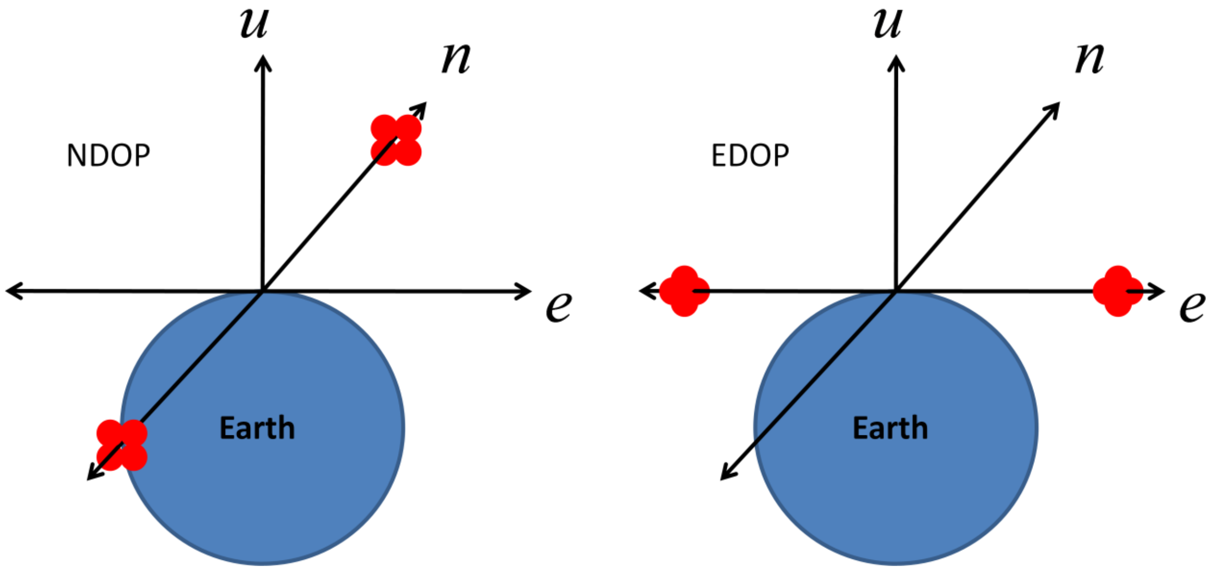


Figure 1.4.4-1

In each figure, there are two groups of four satellites located near each end of the axis in the direction of interest. These configurations are not realistically achievable in GPS due to the real world satellite orbits.

Even in theory these satellites cannot all be exactly on one axis. If they were, all the points would lie on one line causing the system to be singular. The other two coordinate vectors would be all zero, the G matrix would not have full rank, and the $G^T G$ matrix would have no inverse. Also, the associated parallelotope would be flat and have zero volume. The closer the points come to the axis, the better the DOP in that direction. It approaches the limit of $1/\sqrt{m}$ but can never reach it without becoming singular. The limit is $1/\sqrt{m}$ because of the lower bounds

$$NDOP \geq \frac{1}{\sqrt{m} \cdot \sigma_{\vec{n}}} \qquad EDOP \geq \frac{1}{\sqrt{m} \cdot \sigma_{\vec{e}}}$$

and because $\sigma_{\vec{n}}, \sigma_{\vec{e}} \to 1$ in their respective configurations.

It should be apparent that none of EDOP, NDOP, or VDOP can be minimized simultaneously because maximizing the variance in one coordinate precludes maximization in any other direction. This is due to the coordinates being normalized in the Euclidian sense. For some other linear regression application besides GPS, this might not be the case. If for example the coordinates where normalized in the ℓ^{∞}-norm, it could be possible to maximize variance in all directions simultaneously while achieving all zero correlations. Such a configuration for $m = 8$ with a user at the origin would be

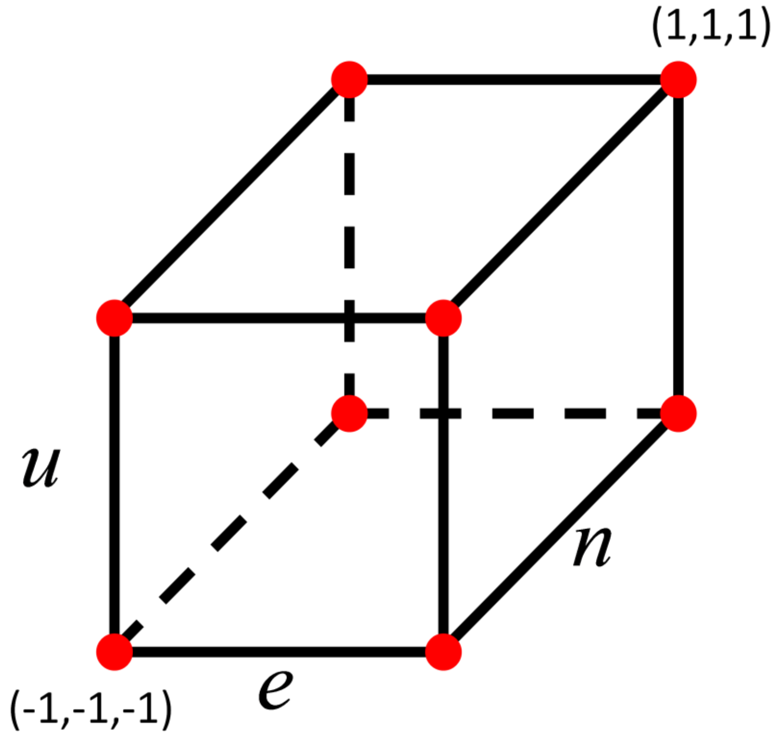


Figure 1.4.4-2

This assumes that negative up coordinates are allowed in the application. In any direction, half of the points are at the minimum and half are at the maximum, so each direction has maximum variance. The correlation is also zero because of symmetry.

1.4.5 TDOP Minimization and a Hyper-tope

The case of TDOP is slightly different. Using the geometry matrix $G = [e\ n\ u\ \mathbf{1}]$ for GPS and equation 1.3.4-3 as before, the DOP is still a ratio of two volumes. However, the coordinate which does not appear in the numerator is the time coordinate represented by the ones column. So, equation 1.3.3-8 is not helpful in this case. Thus the vectors in the equation below are not centered.

$$TDOP = \sqrt{\Sigma_{T,T}} = \sqrt{\frac{|M_{T,T}|}{|H|}} = \frac{Vol(e,n,u)}{Vol(e,n,u,\mathbf{1})} \qquad \text{Eq. 1.4.5-1}$$

This leaves a 3D parallelpiped in the numerator and a 4D hyper-parallelotope in the denominator.   The parallelelpiped is represented in the figure below abstractly as a 2D parallelotope. The fourth dimension is represented with a single edge for the ones vector.

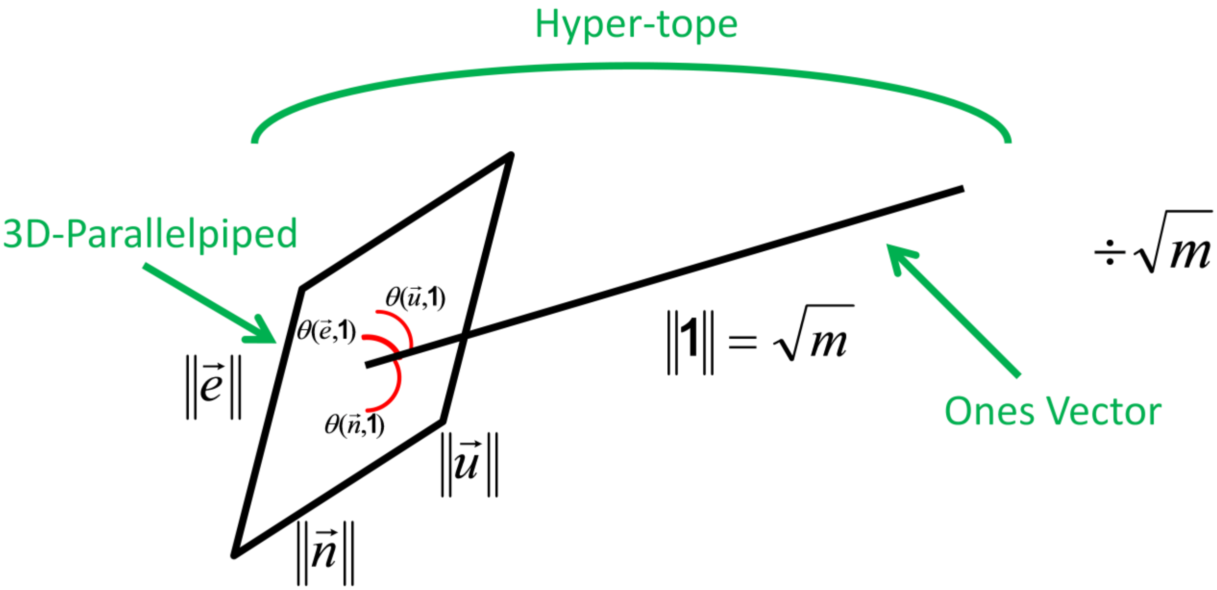

Hyper-tope
3D-Parallelpiped
$\theta(\vec{u},\mathbf{1})$
$\theta(\vec{e},\mathbf{1})$
$\theta(\vec{n},\mathbf{1})$
$\|\vec{e}\|$
$\|\vec{u}\|$
$\|\vec{n}\|$
$\|\mathbf{1}\| = \sqrt{m}$
$\div\sqrt{m}$
Ones Vector

Figure 1.4.5-1

Since the vectors in equation 1.4.5-1 are not centered, neither Lemma 1.3.6-1 or 1.3.6-2 are used. So the vector variances and correlations are not used for the edge lengths and interior angles here.

The goal is still to minimize the ratio of the face to the total parallelotope. The edge sizes $\|\vec{e}\|$, $\|\vec{n}\|$, and $\|\vec{u}\|$ will not have any effect on the ratio since these terms effect the numerator and denominator equally. The final edge length $\|\mathbf{1}\| = \sqrt{m}$ is fixed since the vector is always ones. So, the ratio is minimized simply when the ones edge is orthogonal to the 3D face made of the e, n and u edges. In this case the angles between the e, n, and u edges have no effect on the ratio in the same way that $\rho_{\hat{e},\hat{n}}$ had no effect on VDOP. The ones edge is orthogonal to the 3D face when $\theta(\vec{e},\mathbf{1})$, $\theta(\vec{n},\mathbf{1})$, and $\theta(\vec{u},\mathbf{1})$ are right angles. It can be seen from section 1.3.3 that this happens exactly when e, n, and u are centered vectors. Thus, TDOP is minimized when the e, n, and u coordinates are centered (have zero mean).

The solution for TDOP extends to linear regression in general. Whenever there is a linear model,

$$y = \beta_n x_n + \beta_{n-1} x_{n-1} + \ldots + \beta_1 x_1 + \beta_0 + \varepsilon$$

a design matrix can be chosen to minimize the variance of the linear regression estimator for the y-intercept β_0 simply by choosing data points that are centered for each x_i.

In practice for GPS, e and n can be centered in many configurations. However, since the u coordinates are bounded in [0, 1], the u vector cannot be centered (for a user on the ground). The minimization for TDOP can still be found by maximizing the angle $\theta(\vec{u},\mathbf{1})$. From the definition of dot product

$$\theta(\vec{x},\mathbf{1}) = \cos^{-1}\left(\frac{<\vec{x},\mathbf{1}>}{\|\vec{x}\| \cdot \|\mathbf{1}\|}\right) \qquad \text{Eq. 1.4.5-2}$$

or

$$\theta(\vec{x},\mathbf{1}) = \cos^{-1}\left(\frac{\sum x_i}{\sqrt{m}\cdot\|\vec{x}\|}\right) \qquad \text{Eq. 1.4.5-3}$$

Equation 1.4.5-3 confirms that the angle is a right angle when the vector $\vec{x}$ is centered. It also shows that $\theta(\vec{x},\mathbf{1})$ is closest to a right angle when $\frac{\sum x_i}{\sqrt{m}\cdot\|\vec{x}\|}$ is closest to zero. For a vector with all non-negative elements, such as u, this is the ℓ_1-norm divided by the ℓ_2-norm. The ℓ_1-norm of a vector is bounded below by the ℓ_2-norm. This implies that:

$$\frac{\sum u_i}{\sqrt{m}\cdot\|\vec{u}\|} \geq \frac{1}{\sqrt{m}}$$

It is easy to see that this bound is achieved when only one element of u is non-zero. This can be thought of as being close to centered. So TDOP for GPS is minimized when exactly one satellite is above elevation zero and the satellites are centered across the e and n coordinates meaning that $\bar{e} = \bar{n} = 0$.

1.4.6 HDOP Optimization

A formula for HDOP follows immediately from equations 1.4.4-1 and 1.4.4-2 with equation 1.2.3-5.

$$HDOP = \frac{\sqrt{\sigma_{\vec{e}}^2 \cdot \left(1 - p_{\vec{e},\vec{u}}^2\right) + \sigma_{\vec{n}}^2 \cdot \left(1 - p_{\vec{n},\vec{u}}^2\right)}}{\sqrt{m} \cdot \sigma_{\vec{n}} \cdot \sigma_{\vec{e}} \cdot \sqrt{1 + 2 p_{\vec{e},\vec{n}} p_{\vec{n},\vec{u}} p_{\vec{e},\vec{u}} - p_{\vec{e},\vec{n}}^2 - p_{\vec{n},\vec{u}}^2 - p_{\vec{e},\vec{u}}^2}} \qquad \text{Eq. 1.4.6-1}$$

This can also be written in terms of volumes.

$$HDOP = \frac{\sqrt{Vol^2(\hat{n},\hat{u}) + Vol^2(\hat{e},\hat{u})}}{Vol(\hat{e},\hat{n},\hat{u})} \qquad \text{Eq. 1.4.6-2}$$

The corresponding parallelotope is shown here.

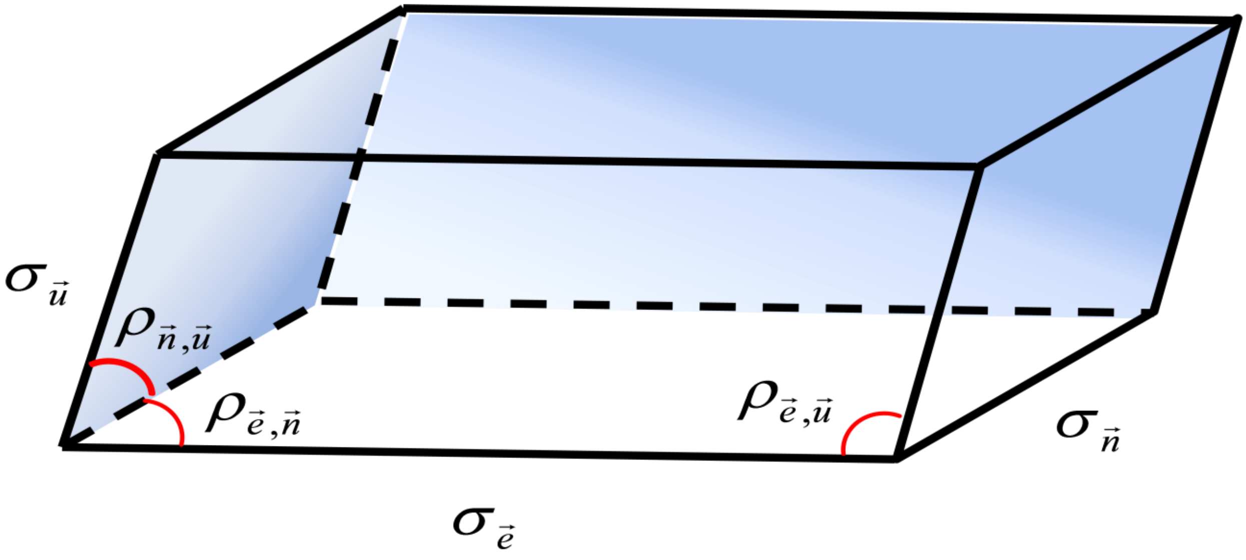


Figure 1.4.6-1

We see here that the numerator is the root sum of squares of the two face volumes. From either equation or the parallelotope, it is clear that $\sigma_{\vec{u}}$ has no bearing on HDOP since it affects both of the faces and the total volume all equally. It is also clear that HDOP is reduced when either $\sigma_{\vec{e}}$ or $\sigma_{\vec{n}}$ are increased. This means that it is beneficial for HDOP to have all satellites at the lowest possible elevation because positioning the satellites in such a way maximizes $\sigma_{\vec{e}}$ and $\sigma_{\vec{n}}$ while ignoring the loss in $\sigma_{\vec{u}}$. As usual, the satellites may not all actually be contained in one plane so that the system remains nonsingular. It is also desirable to have low correlations between each coordinate vector. When the correlations are all zero, the lower bound for HDOP is achieved.

$$HDOP \geq \frac{\sqrt{\sigma_{\vec{e}}^2 + \sigma_{\vec{n}}^2}}{\sqrt{m} \cdot \sigma_{\vec{n}} \cdot \sigma_{\vec{e}}} \tag{Eq. 1.4.6-3}$$

This means that the lowest HDOP occurs when all coordinate vectors are uncorrelated, the satellites are at lowest allowed elevation, and the satellites are spread widely across the east and north directions such that $\frac{\sqrt{\sigma_{\vec{e}}^2 + \sigma_{\vec{n}}^2}}{\sqrt{m} \cdot \sigma_{\vec{n}} \cdot \sigma_{\vec{e}}}$ is minimized.

1.4.7 PDOP & GDOP

PDOP and GDOP can also be written in terms of volumes. The centered forms are not used for GDOP because it contains TDOP.

$$PDOP = \frac{\sqrt{Vol^2(\hat{n},\hat{u}) + Vol^2(\hat{e},\hat{u}) + Vol^2(\hat{e},\hat{n})}}{Vol(\hat{e},\hat{n},\hat{u})} \tag{Eq. 1.4.7-1}$$

$$GDOP = \frac{\sqrt{Vol^2(\vec{n},\vec{u},\mathbf{1}) + Vol^2(\vec{e},\vec{u},\mathbf{1}) + Vol^2(\vec{e},\vec{n},\mathbf{1}) + Vol^2(\vec{e},\vec{n},\vec{u})}}{Vol(\vec{e},\vec{n},\vec{u},\mathbf{1})} \tag{Eq. 1.4.7-2}$$

The denominator of PDOP is the volume of a 3D parallelotope made of centered vectors. The numerator is the root sum of squares of all the face volumes, similar to a surface area. GDOP is very similar in that the numerator is the root some of squares of all the face volumes, but the denominator is a 4D parallelotope volume.

In both cases, optimization requires minimizing the ratio between a root sum of squares surface area (or volume) and the volume of the higher dimensional parallelotope. The surface volume to total volume ratio will grow large if any edge length or interior angle is reduced. For PDOP, which is in centered form, this means that no standard deviation in any coordinate should be low and no two coordinate vectors should have high correlation. For GDOP, it means that no coordinate vector should have low norm and none of the vectors should have a small angle between them. As pointed out in the discussion of TDOP, since one of vectors involved is $\mathbf{1}$, the other vectors should be centered. So, for optimal PDOP and GDOP, there should be a wide spread of satellites in every direction, which is not lopsided in any direction, and with low correlation between each coordinate vector.

1.5 Worst Case Scenarios for DOP

It is useful to also consider what the worst case scenarios are for DOP in addition to the optimal scenarios. Although a given DOP with a set number of satellites has a theoretical minimal bound beyond which cannot be improved, any DOP can approach infinity under certain circumstances. There are no upper bounds. This section will investigate which satellite geometries will cause DOPs to become unbounded.

1.5.1 Known Worst Case Scenarios

In Farrel and Barth (1998) and Farrel (2008), worst case scenarios for DOP are described as occurring when the geometry matrix has columns which are linearly dependent. Linear dependence of the columns of G causes G to be less than full rank and $G^T G$ to be singular. A singular $G^T G$ matrix makes DOP terms incalculable by inversion of the matrix. For practical purposes, the condition of the matrix may be used to determine when the matrix is nearly singular and as a warning for bad DOP. This is a conservative test since no DOP can approach infinity without this happening. However, this does not tell the entire story about worst case scenarios. It is entirely possible, at least in theory, to approach linear dependence in the G matrix without every DOP growing towards infinity or even getting large. There are two specific ways in which a DOP term can approach infinity. These will be discussed in the next two sections. The first is technically a special case of the second, but is worth considering separately.

1.5.2 Zero Variance Singularities

One way for a DOP term to grow towards infinity is for one of the coordinate column vectors to approach zero variance. In GPS this would be zero variance in the e, n, or u coordinates. However, approaching zero variance in one coordinate does not necessarily imply that each DOP will grow. Consider for simplicity the case of VDOP. Recall that in the VDOP formula given as equation 1.4.3-2, there are no σ_e or σ_n terms. So the variances in the e and n coordinates have no bearing on VDOP. So it should be possible to have σ_e and σ_n approach zero without VDOP approaching infinity. Consider the VDOP parallelotope in Figure 1.4.2-2 and remember that VDOP is the area of the 2D face over the volume of the 3D parallelotope. As discussed in 1.4.2, changing the edge lengths σ_e and σ_n has no effect on the DOP since both the area and volume are affected equally. However, the shortening of the edge σ_u towards zero will cause the ratio to increase towards infinity. So VDOP has only one critical coordinate in which approaching variance zero causes a worst case DOP scenario.

In such a scenario, many equivalent things are taking place. The $\vec{u}$ column of G is approaching a constant value which causes it to be linearly dependent with the **1** column vector. G is becoming less than full rank. $G^T G$ is becoming singular. The associated parallelotope is becoming flat in the σ_u direction. Finally, the ratio of the 2D area to the 3D volume is approaching infinity.

The behavior of worst case scenarios for EDOP and NDOP are exactly the same except that the critical coordinates are e and n respectively. TDOP has no such scenario since the time coordinate is represented by the constant **1** vector which already has zero variance.

1.5.3 Total Correlation Singularities

Total correlation singularities occur when two of the coordinate vectors approach a correlation of 1 or -1. In actuality, this is what is happening in the zero variance singularities above since reaching zero variance is equivalent to being totally correlated to the **1** vector. However, thinking of zero variance singularities is useful when using the centered form of a parallelotope as was done with VDOP, EDOP, and NDOP.

As two coordinate vectors become highly correlated, the angle in the parallelotope between those two vectors decreases towards zero. This causes the volume of the parallelotope to approach zero as the parallelotope becomes highly skewed. However, this also will not cause a worst case scenario in every situation. Consider VDOP again and suppose that $p_{\vec{n},\vec{u}} = p_{\vec{e},\vec{u}} = 0$. This means that the σ_u edge of the parallelotope is orthogonal to the 2D face made up by the σ_e and σ_n edges. If this is the case, then reducing the angle represented by $\rho_{\hat{e},\hat{n}}$ will reduce the 2D area and the 3D volume by the same factor. So the ratio of the two, and thus VDOP, are unchanged as was discussed in section 1.4.2. If it is not the case that $p_{\vec{n},\vec{u}} = p_{\vec{e},\vec{u}} = 0$, then $\rho_{\hat{e},\hat{n}}$ will have an effect. Still, this demonstrates that linear dependence will only cause a worst case scenario under most (but not all) circumstances.

All that remains then, is to understand the singularities for DOPs made of multiple terms such as HDOP, PDOP and GDOP. Since each of these DOPs is a root sum of squares of two or more single term DOPs, they will approach infinity if and only if one or more of the single term DOPs does. For example, HDOP will approach infinity if either EDOP or NDOP does. PDOP will approach infinity if HDOP or VDOP does. GDOP will approach infinity if any other DOP does.

1.6 Satellite Subset Selection

Often it is impossible for a device to actually use all the satellites in view to make a DOP calculation. There must be an individual channel for each satellite and a particular device may have many less channels than there are satellites in view at any time. The problem of satellite subset selection is to develop methods of choosing which satellites should be included in the DOP calculations at a given time and which should be ignored. Understanding the behavior of DOP is critical to this process. The purpose of this section is to use the results of the previous sections on best and worst case DOP to form some basic ideas about satellite selection techniques. A simple very fast algorithm is presented and evaluated using standard GPS constellation data.

1.6.1 The Goals of Satellite Subset Selection

For purposes here, the goal of subset selection will be to find methods to quickly choose a subset with a given number of satellites which will produce low values for some particular DOP(s) whether they are single term such as VDOP or a multiple term such as GDOP. Naturally, it is possible given any geometry matrix to calculate the DOP values for every subset of a given size and to choose the subset with the minimum value. However, for a large number of satellites in view, such an algorithm would become intensive and efficient algorithms are necessary in subset selection for optimizing DOP (Farrel & Barth, 1998).

Before going on to develop such algorithms, it will be useful to consider what type of results should be expected of a good selection algorithm. It is important to note that theoretically (i.e. when the measurements errors are uncorrelated and have the same variance), a subset of satellites will never outperform the complete set. This is because the DOP terms are proportional to standard deviations of linear regression estimators which are best linear unbiased estimators (BLUE). They are best in the sense that they have lowest variance. Since a subset of satellites produces another linear unbiased estimator, it cannot be better in terms of having lower variance (or DOP). So then it must be decided what a selection algorithm can be expected to accomplish in terms of DOP values.

In an abstract sense, any given satellite geometry can be considered to have good behavior or poor behavior. As seen in the various DOP formulas developed in section 1.4 and by the parallelotope analysis, good behavior is generally to have high variance in the critical coordinates and low correlations between the coordinate vectors. As discussed in section 1.5, bad behavior can result from low variances or high correlations. If a total-in-view satellite set has good behavior in that it is balanced in each direction and well spread out across the sky, it may be that all or most of the satellites are in desirable locations. When forced to choose a smaller subset, it is likely that some desirable satellites will have to be removed and the resulting subset will be less well behaved than the total set. On the other hand, it is possible for the total set to have bad behavior if it is lopsided in a coordinate or the satellites are not well spread out. In this case there may be satellites which are not contributing much to lower DOP values because they are causing a higher correlation or lower variance in some direction. It is still impossible for a subset to attain a better DOP value than the total set. However if some undesirable satellites are removed the subset may be better behaved than the original set.

A simple idea for measuring the behavior of a subset versus that of the total set is to normalize the DOP formula by the number of satellites. Consider any DOP formula such as 1.4.3-2. There is always a $\sqrt{m}$ term in the denominator where m is the number of satellites. Essentially, this factor is what makes it impossible for a subset of size $m_s < m$ to outperform the original set. Rescaling the DOP of the original set by multiplying by $\frac{\sqrt{m}}{\sqrt{m_s}}$ allows the behavior of the subset to be measured against the behavior of the original set as if they had the same number of satellites. An algorithm for subset selection could be evaluated on the basis of how well the resulting subsets compete with rescaled DOP factors.

1.6.2 Current Algorithms and Scalability Problems

Currently, a method used for subset selection is to choose the highest elevation satellites available. This is done in part because very low elevation satellites tend to suffer from larger pseudo-range errors due to multipath problems and physical obstructions. Also, the algorithm is very efficient since it only requires sorting the satellites in view by their elevations. The order of operations of such a sort is $m\log_2(m)$. However, in terms of DOP this algorithm is very poor. Choosing the highest elevation satellites tends to choose a clump of satellites that are near each other, thus minimizing the variance in each coordinate. From section 1.5.2 this is known to lead to a worst case scenario for DOP. Furthermore, as many more satellites become available due to the use of Galileo and GLONASS in combination with GPS, the algorithm will be more successful at choosing a tight clump of higher elevation satellites given that the number of channels (subset size) remains the same for a receiver. This means that with the addition of these satellites, a receiver using this algorithm will not benefit from improved DOPs, but instead may suffer from much worse DOP values. Obviously, it is highly undesirable for the addition of more equipment into a system to result in worse performance. So it is clear that different algorithms are needed.

1.6.3 A Fast and Scalable Satellite Selection Algorithm

An algorithm is given here which does not have the scalability problems discussed in 1.6.2 and which has a low order of operations. The steps are listed here for a very general form of the algorithm followed by comments on each step and possible design decisions.

Sky Slice Algorithm:

1) Slice the sky into mutually exclusive regions that together span the entire sky.
2) Count the number of satellites in each region of the sky.
3) Successively remove satellites from the most populated region until the desired number of satellites remains.

Step 1) The method of dividing up the sky is critical to the performance of this algorithm and it can be done in various ways. Two simple ideas would be to slice the sky by degree of elevation or by degree of azimuth. However, these are not likely to yield good results because simply spreading the satellites over all elevations or all azimuths will not optimize any DOP. As seen in section 1.4, a good method should try to maximize the variance of the coordinates while minimizing the correlations between them. One option aimed at accomplishing this is to separate the sky into eight equally sized rectangular regions defined by slicing with the planes $e = 0$, $n = 0$, and $u = 1/2$. This is especially convenient if the subset size is $m_s = 8$. In that case an ideal subset selection would contain exactly one satellite in each region. Such a subset is likely to have high variance in each direction with low correlations since there is a symmetry in every direction across the center point at $(e, n, u) = (0,0,1/2)$.

In general the number of regions does not need to equal the subset size. For example, the goal could be to get two satellites in each region. The regions can also be designed in a way aimed at optimizing a particular DOP instead of giving equal consideration to all directions as the above method does.

Step 2) Given that the regions of the sky have been decided, it should be easy to determine how many satellites are in each region at a given time. Since there are m satellites at the start, this step has order of operations m.

Step 3) In order to successively remove satellites from the most populated regions, the regions must first be sorted by satellite count. Assuming there are R regions, this has order of operations $R\log_2(R)$. After each satellite is removed, it is not necessary to perform an entirely new sort. Once the regions are ordered from highest to lowest population, the process is to remove one from the most populated region and then determine if that region must be moved down the ordered list by comparing the new count to the next highest population. The region is moved down the list until the region immediately below it does not have a higher satellite count. At most, there can be R-1 moves from the top of the list to the bottom of the list. Since satellites are removed $m-m_s$ times, the order of operations here is $(m-m_s)\cdot(R-1)$. So the total order of operations for this step is $R\log_2(R)+(m-m_s)\cdot(R-1)$.

An important note about this step is that the satellites can be removed from each region in a clever way instead of by random. As an example, consider the eight rectangular regions discussed above. If VDOP is considered important for the application, a good choice might be to always remove the lowest elevation satellite in any of the rectangles above $u=1/2$, and to always remove the highest elevation satellite in any rectangle below $u=1/2$. This will increase the chances of having some very high and very low elevation satellites without many at mid elevations. This helps cause $\sigma_{\vec{u}}$ to be large and VDOP to be small. Many such considerations could be made depending on the goal of the particular application.

It is also possible to give weights to each region so that the optimum is not when every region has an equal number of satellites. In 1.4 some DOPs are shown to be optimal when there are satellites located near each other. So it could be desirable to have one region with four satellites and then four other regions with one satellite each. In this case, satellites are removed based on how high each region count is above the desired amount for that region instead of removing them based simply on maximum count.

This algorithm avoids scalability problems because a larger number of satellites cannot make it more difficult to balance the number of satellites chosen in each region. In fact, a higher number of satellites in the sky will make it more likely that an optimal balance can be achieved. There also are a variety of ways in which it can be optimized for specific applications as discussed above. Finally, the total order of operations found by adding the operations for step 2 and step 3 is:

$$Operations = m + R\log_2(R) + (m - m_s)\cdot(R-1) \qquad \text{Eq. 1.6.3-1}$$

or

$$Operations = Rm + R\log_2(R) - (R-1)m_s \qquad \text{Eq. 1.6.3-2}$$

This should typically be less than Rm which may be slightly more or less than the $m\log_2(m)$ order of operations for choosing the highest elevation satellites. So the efficiency is approximately the same between the two algorithms.

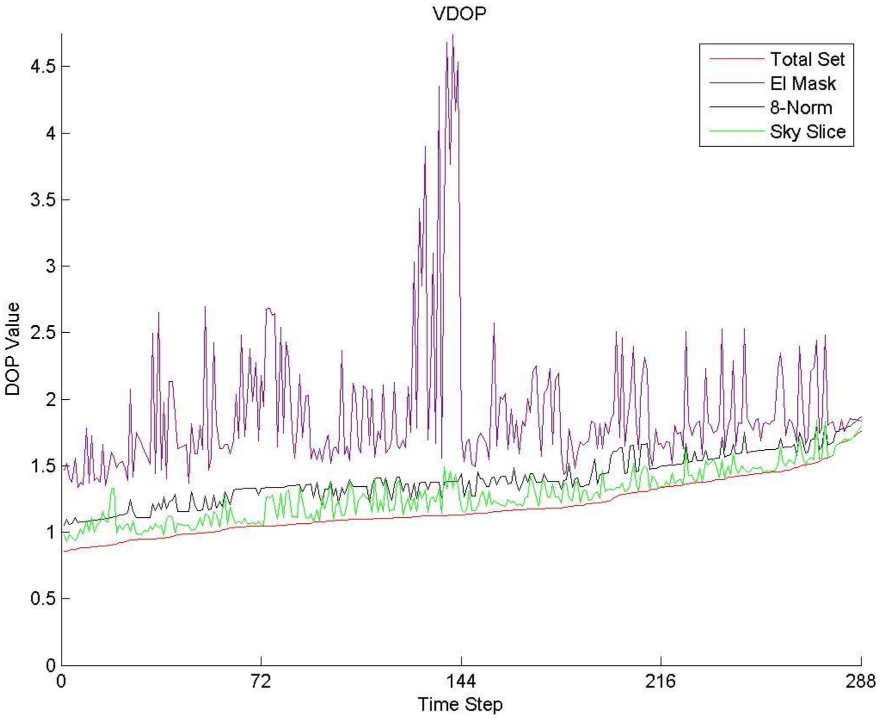
1.6.4 Sky Slicing Algorithm Performance

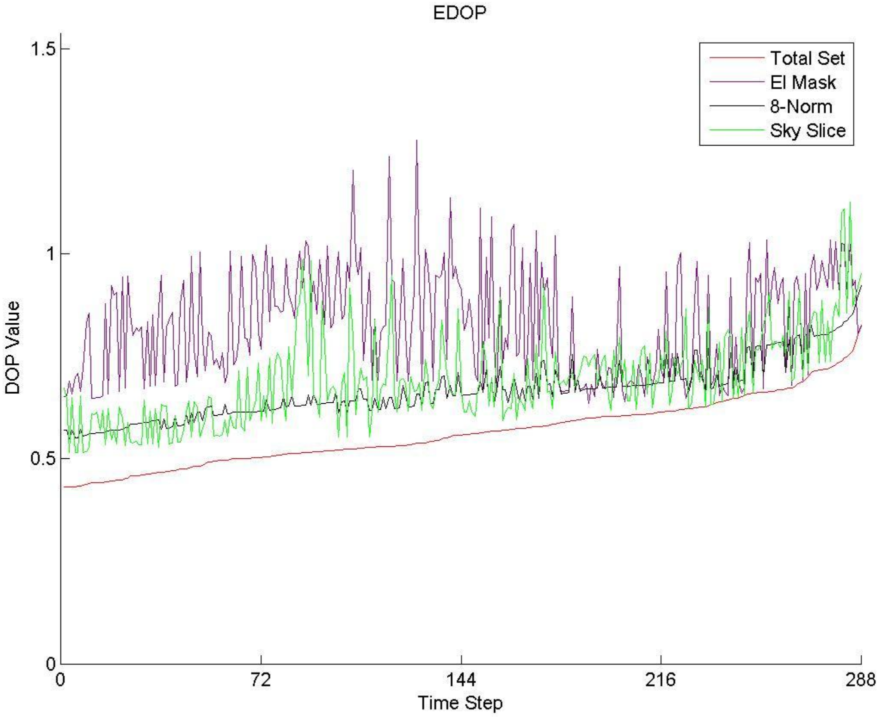
The algorithm in the previous section was run on GPS constellation data for the day of March 12 2006. Since each satellite repeats its orbit twice every sidereal day, data for one day could be expected to closely representative of any other day. A user position was taken to be at 0 latitude and -90 degrees longitude. On this run, the number of satellites in view of the user ranges from 9 to 14 and has a mean of 11.35. The subset size is fixed at $m_s = 8$. The data has a 5 minute decimation time, so there are 288 geometries for the day. Smaller decimation could be used, but there will not usually be much difference between two DOP values if the geometry has not had time to change significantly.

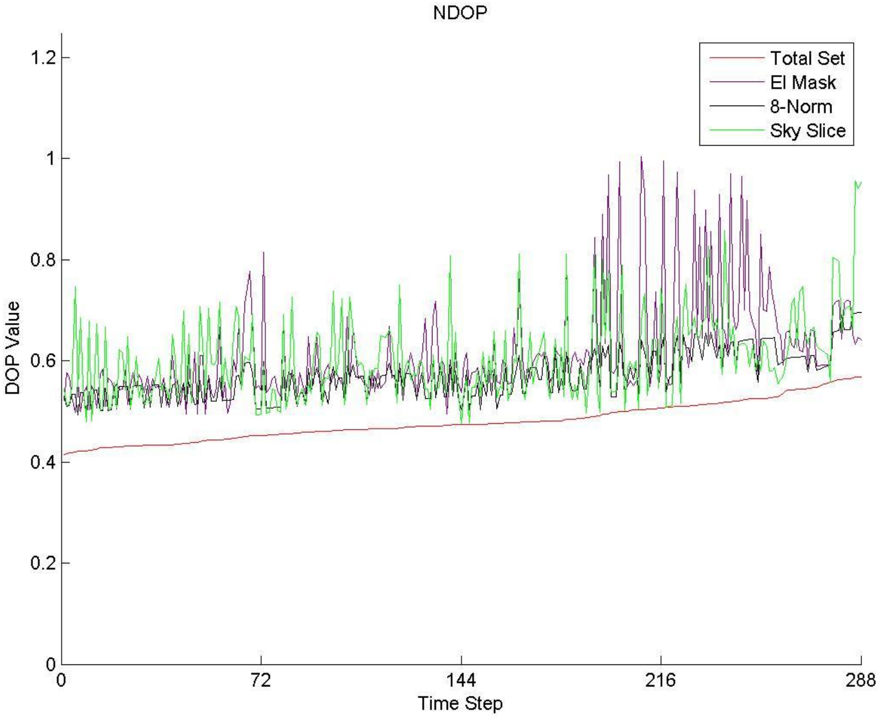
For these runs, the sky is sliced into eight rectangular regions as suggested in the previous section. The satellites are removed from the bottom of the upper regions and from the top of the lower regions.

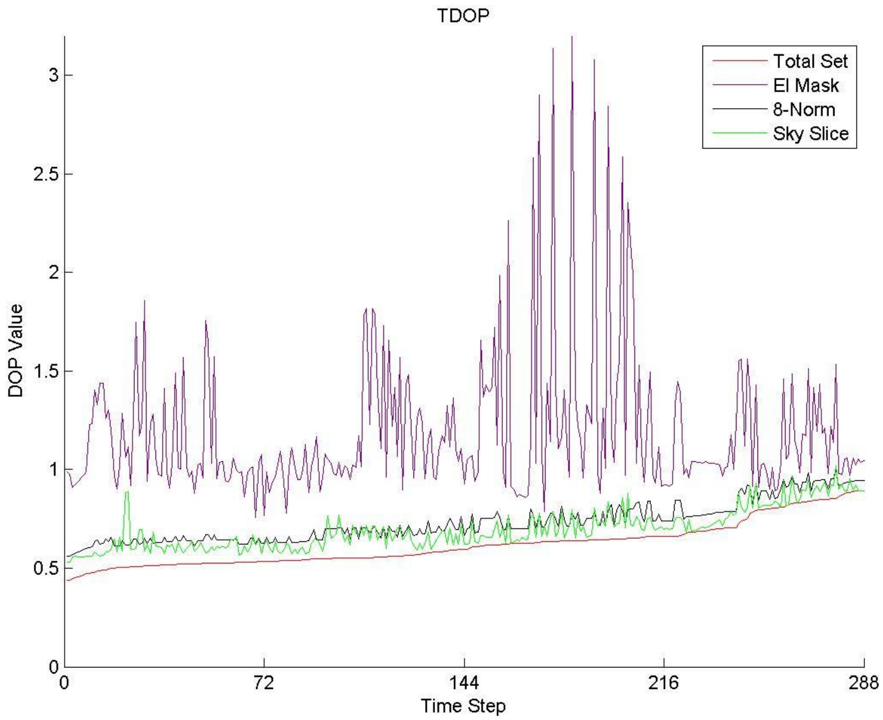
Below there is a figure for each of the seven DOPs investigated in section 1.4. There are four separate data lines on each plot. The DOP of the total in view satellite set (red), the DOP value of the total set normalized to a size of eight satellites as in section 6.1 (black), the DOP of the subset selected by taking the 8 highest elevation satellites (magenta), and the DOP of the subset selected using the sky slicing algorithm (green). The time axis for all of these is sorted by the DOP value of the total set. This causes the total set data to be non-decreasing and makes the plots more readable.

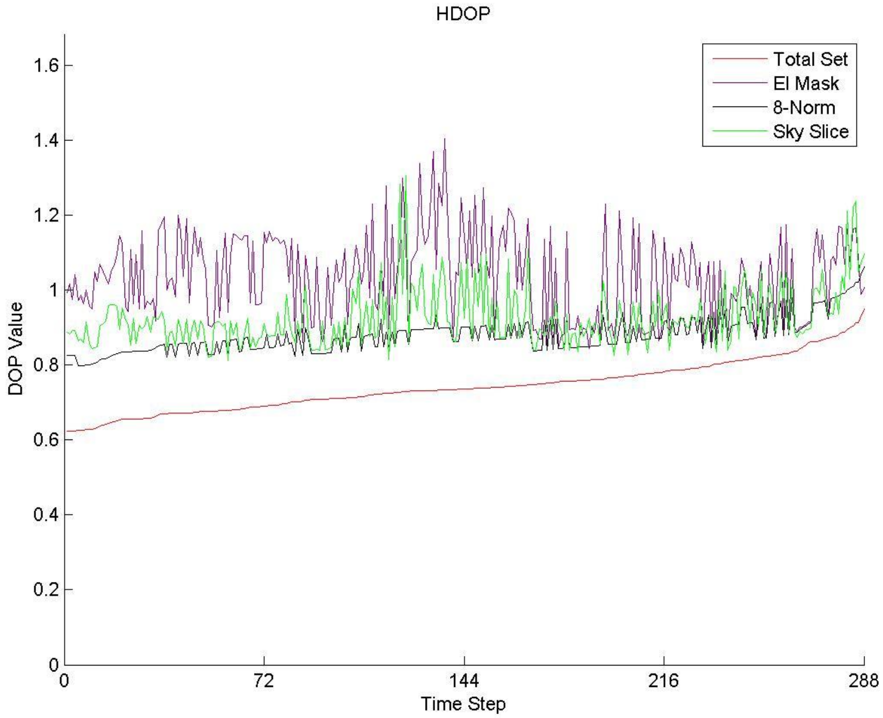
As discussed in section 6.1, it is impossible for any subset to outperform the total set. So the total set data (red) must always be below all the other sets. Also, either of the subsets may be considered to be performing well when they are close to the normalized data (black) as discussed in section 1.6.1. When they approach the total set data, they are performing extremely well.

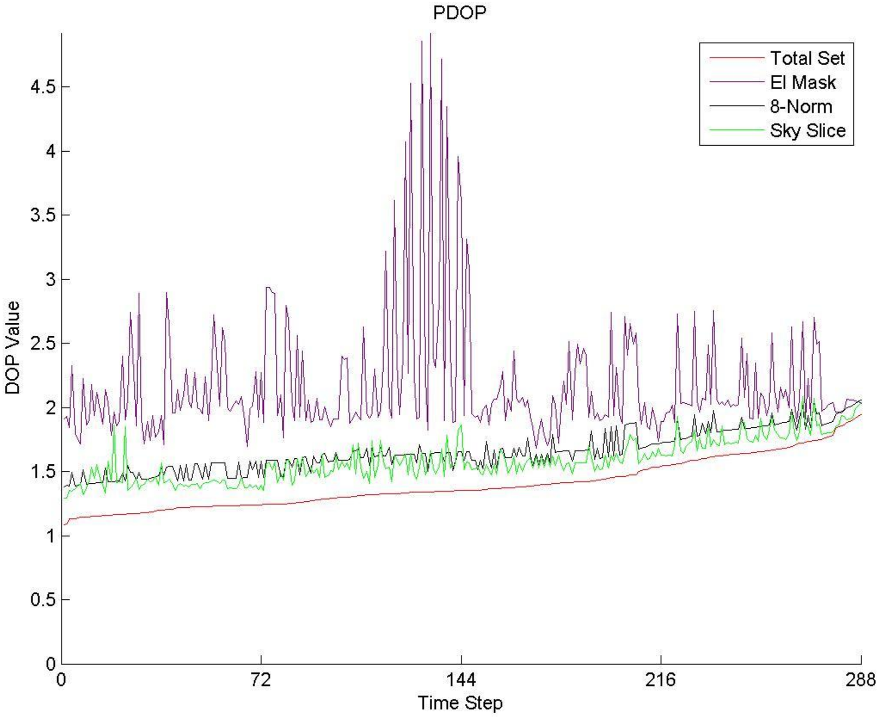


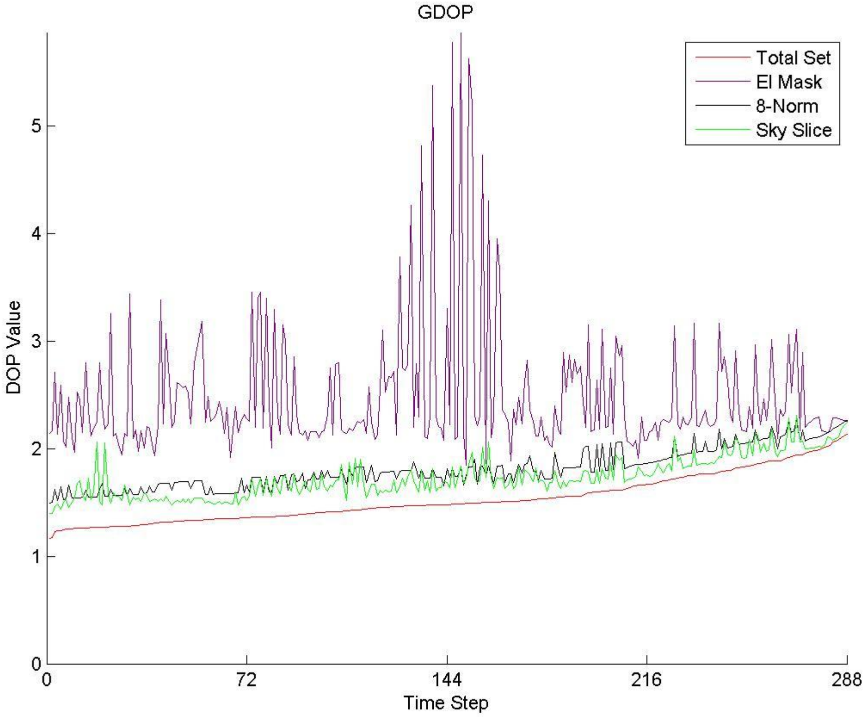

EDOP
Total Set
El Mask
8-Norm
Sky Slice
1.5
1
0.5
0
DOP Value
0
72
144
216
288
Time Step


NDOP
Total Set
El Mask
8-Norm
Sky Slice
1.2
1
0.8
0.6
0.4
0.2
0
DOP Value
0
72
144
216
288
Time Step


TDOP
Total Set
El Mask
8-Norm
Sky Slice
DOP Value
0
0.5
1
1.5
2
2.5
3
0
72
144
216
288
Time Step


HDOP
Total Set
El Mask
8-Norm
Sky Slice
DOP Value
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
0
72
144
216
288
Time Step


PDOP
DOP Value
Time Step
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
0
72
144
216
288
Total Set
El Mask
8-Norm
Sky Slice


GDOP
DOP Value
Time Step
0
1
2
3
4
5
0
72
144
216
288
Total Set
El Mask
8-Norm
Sky Slice

It is clear that the sky slicing algorithm (green) greatly outperforms the highest elevation algorithm (magenta) in VDOP, TDOP, HDOP, PDOP, and GDOP. In each of these cases except HDOP, it is usually below the normalized data line and is often close to the total set line showing that it is performing very well. This particular realization of the sky slicing algorithm is optimized well for VDOP and not HDOP, so these results are expected. A similar realization could easily be optimized for HDOP but would not perform as well for VDOP. It is plain to see that the sky slicing algorithm is superior to the highest elevation algorithm in any case.

1.7 Conclusion

In the preceding sections, a method of conceptualizing DOP was formulated. This parallelotope based geometric interpretation allows a useful understanding of the behavior of DOP regardless of the number of data points in the regression model. These concepts were used to arrive at exact formulas for various DOPs and to find many optimal and worst case scenarios for DOP in GPS. Finally, these results facilitated the creation of an algorithm for satellite subset selection with a low order of operations that performs very well.

This understanding also allows other possible advantages. There is a large variety of known properties for parallelotopes which may prove useful in continued work (Jones 2011; Gover & Krikorian, 2010; Deza & Grishukhin, 2003; Fallat & Johnson, 2000; Leng & Zhang 1998; Schnell, 1994; Ben-Israel, 1992; Miao & Israel, 1991). Other continued work might include an extension for weighted least squares regression, and applications to systems other than GPS.

2.1 Introduction to Cycle-Slips

Accurate positioning of a GPS user requires accurate range measurements to the satellites in view. Since the atmosphere causes delays in the signals used to determine range measurements, the user must have information about these delays to calculate an accurate position. The more challenging and critical of these delays to estimate is that caused by the ionosphere. A cycle-slip is a type of ambiguity that is often found in carrier phase delay data. These cycle-slips must be removed from carrier phase data in order to properly estimate ionosphere delays.

Much research has been done recently on methods of cycle-slip repair for applications in which accurate delay data is needed in real time (Banville & Langley, 2010; Marujao & Mendes, 2007; Teunissen, 2003a; Teunissen, 2003b). Here, the goal instead will be to repair data as a post processing analysis of receiver independent exchange format (RINEX) files. This allows the use of future data in addition to past data to detect and repair cycle-slips. Models will be created to describe the behaviors of the data when cycle-slips are present and when they are not. A test for cycle-slips will be developed which determines which of the models is the most likely given the observed data. Finally, a software tool is created that uses this method to repair RINEX files.

2.2 Pseudoranging and Cycle-Slips

In this section, the basic concepts of satellite pseudoranging are discussed. This includes a simple explanation of how range and delay of a satellite to a receiver are determined. It also includes a description of what cycle-slips are, why they come about, and what difficulties they pose.

2.2.1 Satellite Pseudo-range and Delay

A pseudorange is an approximation of the range between a satellite and receiver which contains some error due to a bias in the receiver clock (Misra & Enge, 2006). Without any delay factor, finding the pseudorange from a satellite to a receiver would be simple. A signal with a time stamp sent from a satellite would travel to the receiver at the speed of light. Multiplying the elapsed time by c (the speed of light) would yield the pseudorange. However, there is a delay in the signal travel time caused by passing through the ionosphere and the troposphere which cause refraction of the signal. A delay in the signal arrival time essentially makes the satellite appear to be farther away. So in order to find the true distance, there must be a means to take the delay into account.

Delays are constantly changing over time due to atmospheric changes, so they must be calculated at each given time. For troposphere delays, there are accurate adjustments based on regional weather conditions. For the more troublesome ionosphere delays, this is done using a collection of precisely surveyed ground stations employing receivers that receive two separate signals sent from every satellite in view. The signals, differentiated by their frequencies, are called L1 and L2 and have frequencies f_1 and f_2 respectively. Each signal produces separate pseudoranges called r_1 and r_2. Each of these is the sum of the ionosphere-free (hereafter abbreviated "iono-free") pseudorange and the ionosphere delay of the corresponding signal.

$$r_1 = r + d_{L1} \qquad \text{Eq. 2.2.1-1}$$

$$r_2 = r + d_{L2} \qquad \text{Eq. 2.2.1-2}$$

The iono-free pseudorange, denoted by r, is the pseudorange that would occur without any signal delay from the ionosphere. d_{L1} and d_{L2} are the delays that appear in the L1 and L2 signals. Because the ionosphere effect on electromagnetic signals is dispersive, the signal delay is frequency-dependent. The relationship between the two delays is:

$$d_{L2} = \frac{f_1^2}{f_2^2} d_{L1} \qquad \text{Eq. 2.2.1-3}$$

Writing equations 2.2.1-1 and 2.2.1-2 as a matrix system with the above substitution yields:

$$\begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & \frac{f_1^2}{f_2^2} \end{bmatrix} \begin{bmatrix} r \\ d_{L1} \end{bmatrix} \qquad \text{Eq. 2.2.1-4}$$

The iono-free pseudorange and delay at L1 can be solved by inverting the matrix.

$$\begin{bmatrix} r \\ d_{L1} \end{bmatrix} = \frac{f_2^2}{f_1^2 - f_2^2} \begin{bmatrix} \frac{f_1^2}{f_2^2} & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \qquad \text{Eq. 2.2.1-5}$$

so that

$$r = \frac{f_1^2}{f_1^2 - f_2^2} r_1 - \frac{f_2^2}{f_1^2 - f_2^2} r_2 \qquad \text{Eq. 2.2.1-6}$$

and

$$d_{L1} = \frac{f_2^2}{f_1^2 - f_2^2} (r_2 - r_1) \qquad \text{Eq. 2.2.1-7}$$

So using a dual-frequency system, the iono-free pseudorange and delay can be calculated (estimated). For any satellite and ground station pair, d_{L1} represents the ionosphere delay caused by the signal passing through the ionosphere on the line of sight between the satellite and station. As an approximation (the so called "thin shell model" of the ionosphere) it can be assumed that this delay occurs all at once at the mean height of the ionosphere layer. This point is called an ionosphere pierce point (IPP). Because there are many satellite / station pairs, there are many IPPs across the sky at any given time. This allows a model for the delays across the ionsphere to be created as a surface. Users of the system, who only receive an L1 signal from each satellite, can use this delay model to correct their L1 pseudorange values and thereby more precisely calculate their positions.

2.2.2 Cycle-Slip Causes and Consequences

The L1 and L2 signals are both sent in two modes leading to two pseudorange measurements, namely the code and carrier phase measurements. The carrier phase measurement is made by detecting and measuring the phase shift of the signal as it travels from satellite to receiver. Because the signal phase angle is ambiguous (i.e. modulo an even multiple of pi), a cycle-slip can arise, causing a corresponding ambiguity in the range measurement. Thus, a cycle-slip is a discontinuity or a jump in the carrier phase signal. These can occur when a ground station temporarily loses signal lock with a satellite. This can be caused by an obstruction between the satellite and ground station antenna, a software failure, or a low signal to noise ratio. Regardless of the cause, when lock is lost and then regained, an unknown number of carrier signal cycles is gained or lost during that time (Kaplan & Hegarty, 2006). The result is a discontinuity in the signal which is proportional to the integer number of cycles lost or gained. Figure 2.2.2-1 is an illustration of a carrier phase delay signal containing two cycle-slips.

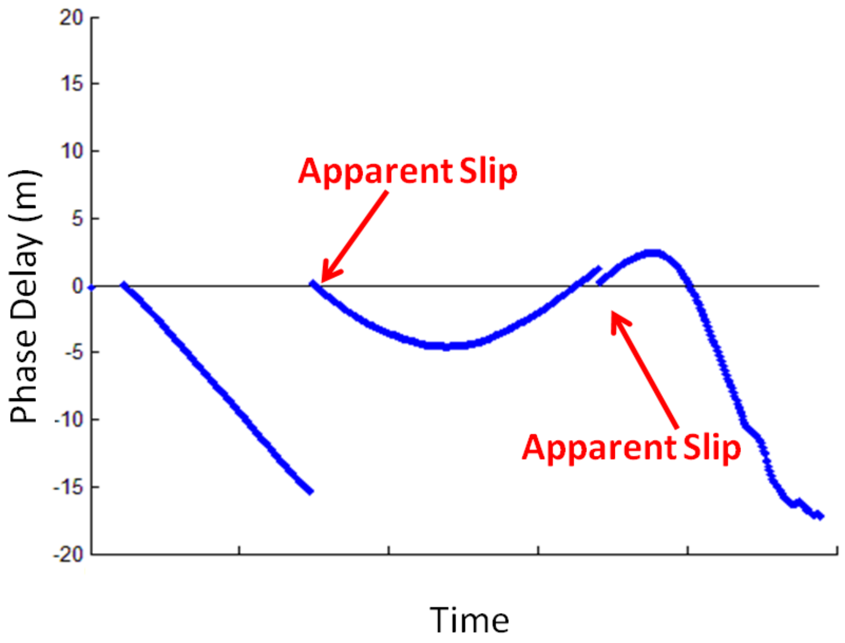


Figure 2.2.2-1

A carrier phase measurement is more precise than a code measurement but is ambiguous and starts at an arbitrary value. Code measurements are unambiguous and do not suffer from cycle-slips. Figure 2.2.2-2 shows the code delay that matches the carrier delay above.

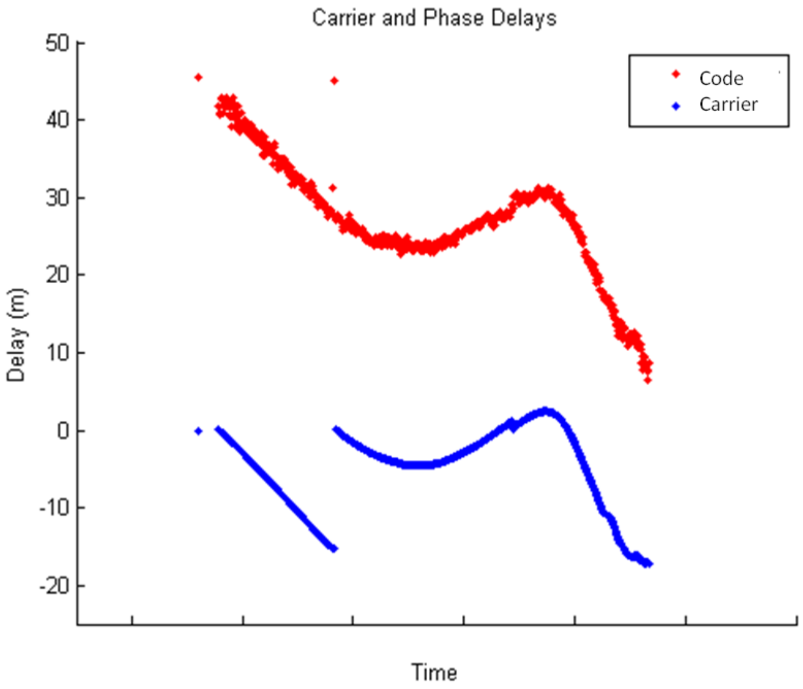


Figure 2.2.2-2

In order to get delay measurements that are both precise and unambiguous, the carrier delay should be leveled (or transposed) onto the code delay. However, this cannot be done effectively with cycle-slips present. Before the data is leveled, the discontinuities in carrier phase caused by cycle-slips must be repaired.

2.3 Cycle-Slip Detection and Repair Using Sequential Hypothesis Testing

In this section, a method will be developed for detecting and repairing cycle-slips in RINEX data in post processing. This means that at any time in a series of delay data, both previous and future data points are used in a test for cycle-slips. Because this method uses future data, it cannot be implemented for real time cycle-slip detection. It is developed for the purpose of removing cycle-slips from data in post processing, but could also be used for near real-time cycle-slip detection with a delay.

2.3.1 Modeling of Delay Data

The method for detecting and repairing cycle-slips developed here assumes that exactly one point (x_0, y_0) is being tested for a slip where x_0 is the time position and y_0 is the signal value. The signal value will be iono-delay, but this method may be applied to detect similar jumps in other time series where the assumptions are the same. It is assumed that only local data are needed to perform the test. This local data will be a neighborhood of $m = 2n$ points about x_0, namely $(X,Y) = \{(x_{-n}, y_{-n}), \ldots, (x_0, y_0), \ldots, (x_{n-1}, y_{n-1})\}$. It is not necessary for the neighborhood size to be even, but this will give some models a symmetry across the possible slip location.

Three models will be presented here for fitting delay data. Each is meant to describe different behavior that is known to occur. The first model, which will be denoted by H_0, describes a neighborhood of data which contains no cycle-slips. This is done with a low order polynomial. Parabolas will be used throughout since they locally fit delay data very well. The inherent assumption in using a polynomial is that the data is smooth in a small neighborhood when no cycle-slips are present. Figure 2.3.1-1 is an illustration for the H_0 model when $n = 5$.

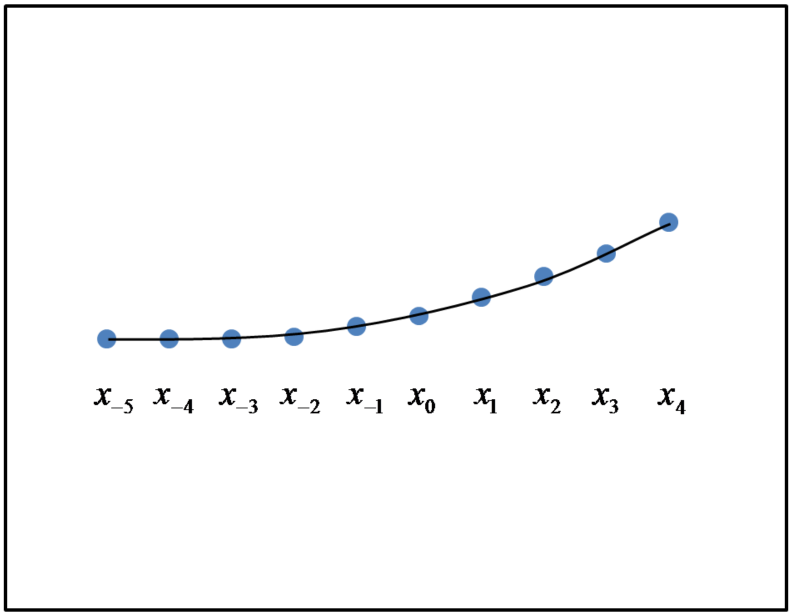


Figure 2.3.1-1

The other two models given here represent two different behaviors when a cycle-slip is present at x_0, the point being tested. These models will be called H_1 and H_2. The corresponding behaviors will be called type 1 and type 2 slips.

The type 1 slip is the cycle-slip which is usually considered. In this case, the delay data is unchanged except that there is a step which occurs at the cycle-slip location. This is modeled with the same low order polynomial with an additional offset term D for $x_t \geq x_0$. This offset term represents a jump size between x_{-1} and x_0. This model is illustrated in Figure 2.3.1-2.

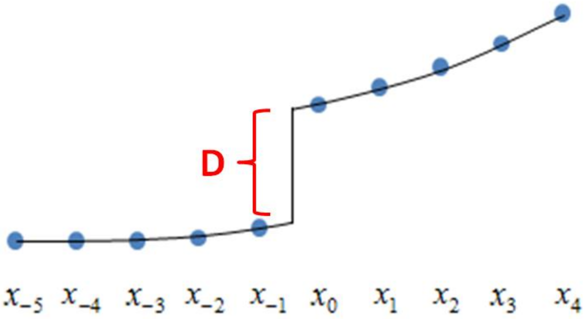


Figure 2.3.1-2

The only other type of slip which will be discussed here explicitly is the type 2 slip. During a type 2 slip, the signal values are smooth throughout the neighborhood except that there is a discontinuity at the slip location and at the location immediately following the slip. This model is added for no other reason than this behavior has been repeatedly observed in delay data calculated from RINEX files. Neither the size nor direction of the two discontinuities are obviously related in any way. It is essentially the same as two consecutive type 1 slips, but since the type 1 slip model will not fit this type of data, another model is required. The type 2 model is again the same low order polynomial, this time with offset terms D_1 and D_2 for $x_t \geq x_0$ and $x_t \geq x_1$ respectively. The type 2 slip model is pictured below.

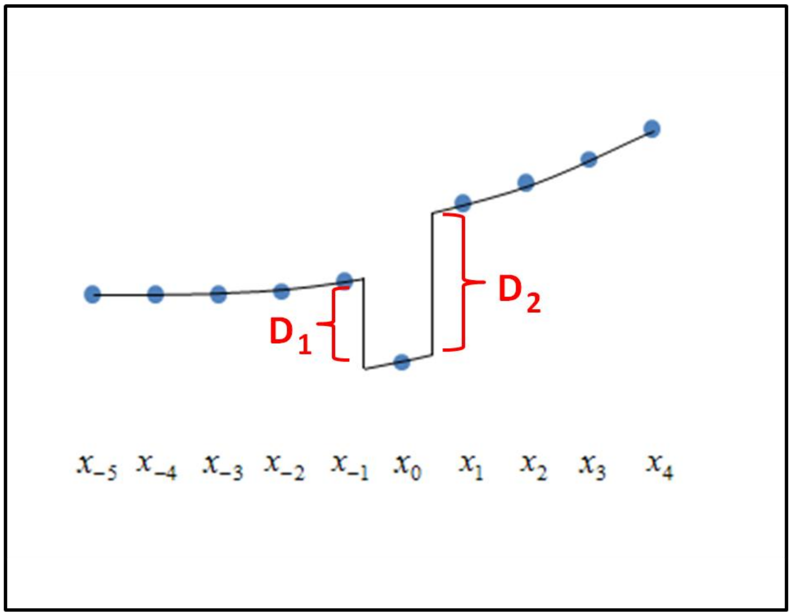

D$_1$
D$_2$
x_{-5} x_{-4} x_{-3} x_{-2} x_{-1} x_0 x_1 x_2 x_3 x_4

Figure 2.3.1-3

These three models describe different behaviors found in delay data. The goal of the next two sections is to describe a method of deciding which of the models is most likely the correct model for a given neighborhood. If H_0 is not the most likely model, then a slip is detected and should be repaired. Repairs are easy to perform by simply removing the estimated offset values in the data.

2.3.2 Fitting Models to a Neighborhood

Each of these models can be fit to a neighborhood of delay data using linear regression. The equations for each, used to form the design matrices, are:

$$H_0: \quad y_t = A + Bx_t + Cx_t^2 \qquad \text{Eq. 2.3.2-1}$$

$$H_1: \quad y_t = A + Bx_t + Cx_t^2 + Du(x_t, x_0) \qquad \text{Eq. 2.3.2-2}$$

$$H_2: \quad y_t = A + Bx_t + Cx_t^2 + D_1u(x_t, x_0) + D_2u(x_t, x_1) \qquad \text{Eq. 2.3.2-3}$$

where u is defined by:

$$u(x_1, x_2) = \begin{cases} 1, & x_1 \geq x_2 \\ 0, & x_1 < x_2 \end{cases} \qquad \text{Eq. 2.3.2-4}$$

So each linear regression has the usual form:

$$\vec{y} = X_k \vec{\beta}_k + \vec{\varepsilon}_k \qquad \text{Eq. 2.3.2-5}$$

Here, $\vec{y}$ is the column vector of $2n$ signal values in the neighborhood, $\vec{\varepsilon}_k$ is the random error vector, $\vec{\beta}_k$ is the parameter vector for the given model, and X_k is the design matrix for the given model. It is assumed that X_k is always of full rank. For the models defined above, the parameter vectors are:

$$\beta_0 = \begin{bmatrix} A \\ B \\ C \end{bmatrix}, \qquad \beta_1 = \begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix}, \qquad \beta_2 = \begin{bmatrix} A \\ B \\ C \\ D_1 \\ D_2 \end{bmatrix} \qquad \text{Eq. 2.3.2-6}$$

The linear regression estimators for the parameter vectors are given by:

$$\vec{b}_k = \left(X_k^T X_k\right)^{-1} X_k^T \vec{y} \qquad \text{Eq. 2.3.2-7}$$

Each $\vec{b}_k$ is the best linear unbiased estimator for the corresponding $\vec{\beta}_k$ when model k is the true underlying model, and defines the fit curve for the k-th model. The sum of squared residuals is of special interest in the following sections. This is the sum of all the squared differences between the fit model and the actual data values. For the k-th model, the sum of squared residuals is written:

$$SS_{res,k} = \vec{e}_k^T \vec{e}_k = \left(\vec{y} - X_k \vec{b}_k\right)^T \left(\vec{y} - X_k \vec{b}_k\right) = \vec{y}^T \left(I - X_k \left(X_k^T X_k\right)^{-1} X_k^T\right)\vec{y} \qquad \text{Eq. 2.3.2-8}$$

where $\vec{e}_k$ is the vector of distances between the fit curve and the data values $\vec{y}$ in the k-th model. This represents the amount of sum of squares accounted for by the random errors, and, when the model is incorrect, lack of fit. Another important term is the regression sum of squares for the k-th model which is written

$$SS_{reg,k} = \vec{y}^T \vec{y} - \vec{e}_k^T \vec{e}_k = SS_{Total} - SS_{res,k} = \vec{y}^T X_k \left(X_k^T X_k\right)^{-1} X_k^T \vec{y} \qquad \text{Eq. 2.3.2-9}$$

This represents the amount of total sum of squares accounted for by the model. It is assumed that the random errors are iid normally distributed for a correct model so that

$$\vec{\varepsilon}_k \sim N(0, \sigma^2 I_{2n\times 2n}) \qquad \text{Eq. 2.3.2-10}$$

which, along with equation 2.3.2-5 implies that

$$y \sim N(X_k \beta, \sigma^2 I_{2nx2n}) \qquad \text{Eq. 2.3.2-11}$$

Based on this assumption, it can be shown that both $SS_{res,k} / \sigma^2$ and $SS_{reg,k} / \sigma^2$ follow chi-squared distributions.

Def (2.3.2-1): Let $\vec{y} \sim N(\mu, I)$ and A be an mxm symmetric idempotent matrix, then $\vec{y}^T A\vec{y}$ follows the chi-squared distribution $\vec{y}^T A\vec{y} \sim \chi^2_{p,\lambda}$ where $p = rank(A)$ is the degrees of freedom and $\lambda = \frac{1}{2}\mu^T A\mu$ is called the non-centrality parameter. A chi-squared distribution is called central and is written χ^2_p if $\lambda = 0$.

Def (2.3.2-2): A matrix B is called idempotent if $B^2 = B$.

It is shown in the appendix that $I - X_k\left(X_k^T X_k\right)^{-1} X_k^T$ and $X_k\left(X_k^T X_k\right)^{-1} X_k^T$ are symmetric and idempotent. It follows that when the model is correct, $SS_{res,k} / \sigma^2$ has a central chi-squared distribution with $2n - p_k$ degrees of freedom where p_k is the rank of X_k:

$$\frac{SS_{res,k}}{\sigma^2} \sim \chi^2_{2n-p_k} \qquad \text{Eq. 2.3.2-12}$$

and that $SS_{reg,k} / \sigma^2$ has a non-central chi-squared distribution with p_k degrees of freedom and non-centrality parameter $\lambda_{reg,k} = \frac{1}{2\sigma^2}\beta^T X_k^T X_k \beta$:

$$\frac{SS_{reg,k}}{\sigma^2} \sim \chi^2_{p_k,\lambda_{reg,k}} \qquad \text{Eq. 2.3.2-13}$$

The non-centrality parameters are calculated by taking

$$\begin{aligned}\lambda_{res,k} &= \frac{1}{2}E(\vec{y}^T)\left(I - X_k\left(X_k^T X_k\right)^{-1} X_k^T\right)E(y)\\ &= \frac{1}{2\sigma^2}\beta^T X_k^T\left(I - X_k\left(X_k^T X_k\right)^{-1} X_k^T\right)X_k\beta\\ &= \frac{1}{2\sigma^2}\beta^T X_k^T X_k\beta - \beta^T X_k^T X_k\left(X_k^T X_k\right)^{-1} X_k^T X_k\beta\\ &= \frac{1}{2\sigma^2}\beta^T X_k^T X_k\beta - \beta^T X_k^T X_k\beta\\ &= 0\end{aligned} \qquad \text{Eq. 2.3.2-14}$$

and

$$\begin{aligned}\lambda_{reg,k} &= \frac{1}{2}E(\vec{y}^T)\left(X_k\left(X_k^T X_k\right)^{-1} X_k^T\right)E(y)\\ &= \frac{1}{2\sigma^2}\beta^T X_k^T\left(X_k\left(X_k^T X_k\right)^{-1} X_k^T\right)X_k\beta\\ &= \frac{1}{2\sigma^2}\beta^T X_k^T X_k\left(X_k^T X_k\right)^{-1} X_k^T X_k\beta\\ &= \frac{1}{2\sigma^2}\beta^T X_k^T X_k\beta\end{aligned} \qquad \text{Eq. 2.3.2-15}$$

2.3.3 F-Tests for Cycle-Slip Detection

Since the models above are embedded, that is the model H_i is embedded within H_j whenever $i < j$, the decision of whether a cycle-slip has occurred or not will be made by performing a sequential hypothesis test. The goal is to create α-level tests that will not only determine if there is a cycle-slip but also what type it is. The tests are based on analysis of variance (ANOVA), meaning that the total sum of squares is broken into components which have meaning with regards to the models' fit of the data. The regression sum of squares and sum of squares residuals are two such components already mentioned in the previous section. The analysis is performed by using the Cochran-Fisher theorem.

Cochran-Fisher Theorem: Let $\vec{y}$ be an $m \times 1$ vector with distribution $N(\vec{\mu}, I)$ such that

$$\vec{y}^T \vec{y} = \sum_{k=1}^{m} \vec{y}^T A_k \vec{y} \,,$$

then the random variables $\vec{y}^T A_k \vec{y}$, k=1,2,…,m, are mutually independent and each has a chi-squared distribution with degrees of freedom $r_k = rank(A_k)$ and non-centrality parameter $\lambda_k = \frac{1}{2} \vec{\mu}^T A_k \vec{\mu}$ if and only if $\sum_{k=1}^{m} rank(A_k) = m$.

Assume that the H_2 model is the overarching model. To reach the goal of creating an α-level threshold, the Cochran-Fisher theorem is applied to the identity

$$\begin{aligned}\frac{\vec{y}^T \vec{y}}{\sigma^2} = & \left(\frac{\vec{y}}{\sigma}\right)^T \left(I - X_2 (X_2^T X_2)^{-1} X_2^T\right) \left(\frac{\vec{y}}{\sigma}\right) \\ & + \left(\frac{\vec{y}}{\sigma}\right)^T \left(X_2 (X_2^T X_2)^{-1} X_2^T - X_1 (X_1^T X_1)^{-1} X_1^T\right) \left(\frac{\vec{y}}{\sigma}\right) \\ & + \left(\frac{\vec{y}}{\sigma}\right)^T \left(X_1 (X_1^T X_1)^{-1} X_1^T - X_0 (X_0^T X_0)^{-1} X_0^T\right) \left(\frac{\vec{y}}{\sigma}\right) \\ & + \left(\frac{\vec{y}}{\sigma}\right)^T \left(X_0 (X_0^T X_0)^{-1} X_0^T\right) \left(\frac{\vec{y}}{\sigma}\right)\end{aligned} \qquad \text{Eq. 2.3.3-1}$$

It is easy to see that this equation holds from cancellation by simply adding the terms on the right hand side. Also, it follows from equation 2.3.2-11 that

$$\frac{\vec{y}}{\sigma} \sim N\left(\frac{1}{\sigma} X_2 \beta, I\right) \qquad \text{Eq. 2.3.3-2}$$

under the assumptions laid out for the cycle-slip problem. So equation 2.3.3-1 has the form of that in the Cochran-Fisher theorem. So in order to see that each of the quadratic forms

$$Q_1 = \left(\frac{\vec{y}}{\sigma}\right)^T \left(X_0 (X_0^T X_0)^{-1} X_0^T\right)\left(\frac{\vec{y}}{\sigma}\right)$$

$$Q_2 = \left(\frac{\vec{y}}{\sigma}\right)^T \left(I - X_2 (X_2^T X_2)^{-1} X_2^T\right)\left(\frac{\vec{y}}{\sigma}\right)$$

$$Q_3 = \left(\frac{\vec{y}}{\sigma}\right)^T \left(X_2 (X_2^T X_2)^{-1} X_2^T - X_1 (X_1^T X_1)^{-1} X_1^T\right)\left(\frac{\vec{y}}{\sigma}\right)$$

$$Q_4 = \left(\frac{\vec{y}}{\sigma}\right)^T \left(X_1 (X_1^T X_1)^{-1} X_1^T - X_0 (X_0^T X_0)^{-1} X_0^T\right)\left(\frac{\vec{y}}{\sigma}\right)$$

are independent and chi-squared distributed, it only needs to be shown that the ranks of the matrices

$$A_1 = X_0 (X_0^T X_0)^{-1} X_0^T$$

$$A_2 = I - X_2 (X_2^T X_2)^{-1} X_2^T$$

$$A_3 = X_2 (X_2^T X_2)^{-1} X_2^T - X_1 (X_1^T X_1)^{-1} X_1^T$$

$$A_4 = X_1(X_1^T X_1)^{-1} X_1^T - X_0(X_0^T X_0)^{-1} X_0^T$$

sum to m. The reason that these matrices have been chosen will become apparent. It happens that each of these matrices is symmetric and idempotent. Proofs of this are included in the appendix. One of the most important properties of idempotent matrices is that their rank is equal to their trace. So, it will suffice to show that the sum of the traces of the three matrices is m. This follows directly from the general property of trace that $\sum_{k=1}^{m} tr(Z_k) = tr\left(\sum_{k=1}^{m} Z_k\right)$.

Applying these properties to the four matrices above shows that

$$\sum_{k=1}^{4} rank(A_k) = \sum_{k=1}^{4} tr(A_k) = tr\left(\sum_{k=1}^{4} A_k\right) = tr(I) = m \qquad \text{Eq. 2.3.3-3}$$

So the condition of the Cochran-Fisher theorem is satisfied. The four quadratic forms are independent and chi-squared distributed. In order to find an α -level threshold for testing, the exact distributions must be known. Under the assumption that all the models are full rank,

$$p_0 = rank(X_0) = rank\left(X_0\left(X_0^T X_0\right)^{-1} X_0^T\right) = 3 \qquad \text{Eq. 2.3.3-4}$$

$$p_1 = rank(X_1) = rank\left(X_1\left(X_1^T X_1\right)^{-1} X_1^T\right) = 4 \qquad \text{Eq. 2.3.3-5}$$

$$p_2 = rank(X_2) = rank\left(X_2\left(X_2^T X_2\right)^{-1} X_2^T\right) = 5 \qquad \text{Eq. 2.3.3-6}$$

Also, because $X_0(X_0^T X_0)^{-1} X_0^T$, $X_1(X_1^T X_1)^{-1} X_1^T$, $X_2(X_2^T X_2)^{-1} X_2^T$, A_3, and A_4 are all idempotent

$$rank(A_2) = rank(I) - rank\left(X_2(X_2^T X_2)^{-1} X_2^T\right) = m - p_2 = m - 5 \qquad \text{Eq. 2.3.5-7}$$

$$rank(A_3) = rank\left(X_2(X_2^T X_2)^{-1} X_2^T\right) - rank\left(X_1(X_1^T X_1)^{-1} X_1^T\right) = p_2 - p_1 = 1 \quad \text{Eq. 2.3.5-8}$$

and

$$rank(A_4) = rank\left(X_1(X_1^T X_1)^{-1} X_1^T\right) - rank\left(X_0(X_0^T X_0)^{-1} X_0^T\right) = p_1 - p_0 = 1 \quad \text{Eq. 2.3.3-9}$$

This implies that the quadratic forms all have chi-squared distributions. It is shown in the appendix that Q_2, Q_3, and Q_4 are each central under the appropriate hypotheses to be tested. So, under those hypotheses,

$$Q_2 \sim \chi^2_{m-5} \quad \text{Eq. 2.3.3-10}$$

$$Q_3 \sim \chi^2_1 \quad \text{Eq. 2.3.3-11}$$

$$Q_4 \sim \chi^2_1 \quad \text{Eq. 2.3.3-12}$$

Since these are independent, they can be used to create F-statistics.

Def (2.3.3-1): Given a non-central chi-squared random variable $Z_1 \sim \chi^2_{r_1,\lambda}$ and a central chi-squared random variable $Z_2 \sim \chi^2_{r_2}$ that are independent of each other, the random variable $F_{r_1,r_2,\lambda} = \frac{Z_1 / r_1}{Z_2 / r_2}$ is non-central F-distributed with degrees of freedom r_1 and r_2 and non-centrality parameter λ. It is called central and written F_{r_1,r_2} if $\lambda = 0$ (the numerator is central chi-squared distributed).

The F-statistics that will be used are:

$$\frac{Q_3/1}{Q_2/(m-5)} = \frac{SS_{reg,2} - SS_{reg,1}}{SS_{res,2}/(m-5)} \sim F_{1,m-5} \qquad \text{Eq. 2.3.3-13}$$

and

$$\frac{Q_4/1}{Q_2/(m-5)} = \frac{SS_{reg,1} - SS_{reg,0}}{SS_{res,2}/(m-5)} \sim F_{1,m-5} \qquad \text{Eq. 2.3.3-14}$$

A large value for equation 2.3.3-13 indicates that the regression model H_2 accounts for much more of the total sum of squares than the H_1 model does alone. In other words, under a null hypothesis that the second offset D_2 is zero, a large value of equation 2.3.3-13 will cause the null hypothesis to be rejected and H_2 accepted as the correct model. If instead the value is small, then H_2 is rejected and equation 2.3.3-14 is used as the next test. This time, the null hypothesis is that the first offset D_1 is zero and (since the H_2 model was previously rejected) $D_2 = 0$. A large value of equation 2.3.3-14 indicates that H_1 accounts for much more of the total sum of squares than the H_0 model alone. So if the value is large, then H_0 is rejected in favor of H_1. Otherwise, the polynomial model with no offsets is accepted. Since each of the tests are based on statistics with central F-distributions, the thresholdsused to accept or reject each hypothesis are given by

$$\frac{SS_{reg,2} - SS_{reg,1}}{SS_{res,2}/(m-5)} > F_{1,m-5}(1-\alpha) \qquad \text{Eq. 2.3.3-15}$$

for the H_2 versus H_1 test and

$$\frac{SS_{reg,1} - SS_{reg,0}}{SS_{res,2} /(m-5)} > F_{1,m-5}(1-\alpha) \qquad \text{Eq. 2.3.3-16}$$

for the H_1 versus H_0 test.

2.3.4 When No Model Fits

The tests developed in the previous section are useful for determining which of the models best describes the data. However, it is possible that none of these models fit well in some given neighborhood. Naturally, since the models are embedded, this is the case if and only if the H_2 model does not fit the neighborhood. This happens for at least two reasons that are common. The first reason is that the data may be very noisy, possibly due to an ionospheric storm. When the data is very noisy, it may be impossible for cycle-slip methods to find cycle-slips. Indeed it may be impossible for a human examination to make the determination. The second reason is that there may be a slip in the neighborhood, but it is not located at (x_0, y_0). The models above are not meant to correctly fit data which has a slip off center.

When the data is very noisy, cycle-slip detections should not be declared and no repair action can be taken. It may be impossible in this case to do anything but ignore the possibility of cycle-slips. When there is a cycle-slip in the neighborhood that is off center, it is possible to use a shrinking window with some minimum size until the H_2 model fits. An even simpler idea is to just use small windows to begin with. The tests defined in the previous sections are effective for neighborhoods as small as $2n = 12$ and possibly smaller.

In any case, there must be a method for determining when none of the models fit because it is entirely possible that the H_1 or H_2 model will comparatively fit some neighborhood much better than the H_0 model even though H_2 is not a good fit. According to the tests in equations 2.3.3-15 and 2.3.3-16, this would cause a cycle-slip detection, but this is a false positive. Since the data is not fit by H_2, it does not actually match either cycle-slip type. The statistic used for testing that H_2 fits the data will be:

$$\frac{m-5}{5}\cdot\frac{SS_{reg,2}-\frac{1}{m}\left(\sum_{i=1}^{m}y_i\right)^2}{SS_{res,2}} \qquad \text{Eq. 2.3.4-1}$$

This is essentially a ratio of the total sum of squares accounted for by the H_2 model given that there is an intercept to the amount accounted for by the random errors. When the model is a correct fit, this ratio should be large. The null hypothesis being tested here is that

$$\vec{\beta}=\begin{bmatrix} A \\ B \\ C \\ D_1 \\ D_2 \end{bmatrix}=\begin{bmatrix} A \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \text{Eq. 2.3.4-2}$$

which corresponds to the idea that H_2 does not account for the date when an intercept is assumed. So under the null hypothesis, when the H_2 model is incorrect, the ratio is expected to be small. The exact distribution of the statistic is found by using the Cochran-Fisher theorem in the same way as the previous section. This time, the quadratic form used is:

$$\begin{aligned}\frac{y^T y}{\sigma^2} &= \frac{y^T}{\sigma}\left(I - X_2\left(X_2^T X_2\right)^{-1} X_2^T\right)\frac{y}{\sigma} \\ &+ \frac{y^T}{\sigma}\left(X_2\left(X_2^T X_2\right)^{-1} X_2^T - \mathbf{1}\left(\mathbf{1}^T\mathbf{1}\right)^{-1}\mathbf{1}^T\right)\frac{y}{\sigma} \\ &+ \frac{y^T}{\sigma}\left(\mathbf{1}\left(\mathbf{1}^T\mathbf{1}\right)^{-1}\mathbf{1}^T\right)\frac{y}{\sigma}\end{aligned} \qquad \text{Eq. 2.3.4-3}$$

where **1** is the $mx1$ submatrix of X_2 that contains a single column of ones representing the intercept. Since the matrices $\mathbf{1}\left(\mathbf{1}^T\mathbf{1}\right)^{-1}\mathbf{1}^T$, $X_2\left(X_2^T X_2\right)^{-1} X_2^T - \mathbf{1}\left(\mathbf{1}^T\mathbf{1}\right)^{-1}\mathbf{1}^T$ and $I - X_2\left(X_2^T X_2\right)^{-1} X_2^T$ are known to be symmetric and idempotent from results in the appendix with ranks 1, 4 and $m-5$ the ranks again sum to m. So again the Cochran-Fisher theorem is satisfied and the two needed quadratic forms

$$\frac{y^T}{\sigma}\left(X_2\left(X_2^T X_2\right)^{-1} X_2^T - \mathbf{1}\left(\mathbf{1}^T\mathbf{1}\right)^{-1}\mathbf{1}^T\right)\frac{y}{\sigma} \qquad \text{Eq. 2.3.4-5}$$

and

$$\frac{SS_{res,2}}{\sigma^2} = \frac{y^T}{\sigma}\left(I - X_2\left(X_2^T X_2\right)^{-1} X_2^T\right)\frac{y}{\sigma} \qquad \text{Eq. 2.3.4-6}$$

are independent and chi-squared distributed. $SS_{res,2}/\sigma^2$ was shown to be central in equation 2.3.2-14. The other quadratic can be rewritten

$$\begin{aligned}
&\frac{y^T}{\sigma}\left(X_2\left(X_2^T X_2\right)^{-1} X_2^T - \mathbf{1}\left(\mathbf{1}^T\mathbf{1}\right)^{-1}\mathbf{1}^T\right)\frac{y}{\sigma} \\
&\quad = \frac{SS_{reg,2}}{\sigma^2} - \frac{y^T}{\sigma}\left(\mathbf{1}\left(\mathbf{1}^T\mathbf{1}\right)^{-1}\mathbf{1}^T\right)\frac{y}{\sigma} \\
&\quad = \frac{SS_{reg,2}}{\sigma^2} - \frac{1}{m}\cdot\frac{y^T}{\sigma}\left(\mathbf{1}\,\mathbf{1}^T\right)\frac{y}{\sigma} \qquad \text{Eq. 2.3.4-5} \\
&\quad = \frac{SS_{reg,2}}{\sigma^2} - \frac{1}{m\sigma^2}\cdot\left(y^T\mathbf{1}\right)\left(\mathbf{1}^T y\right) \\
&\quad = \frac{SS_{reg,2} - \frac{1}{m}\left(\sum_{i=1}^{m} y_i\right)^2}{\sigma^2}
\end{aligned}$$

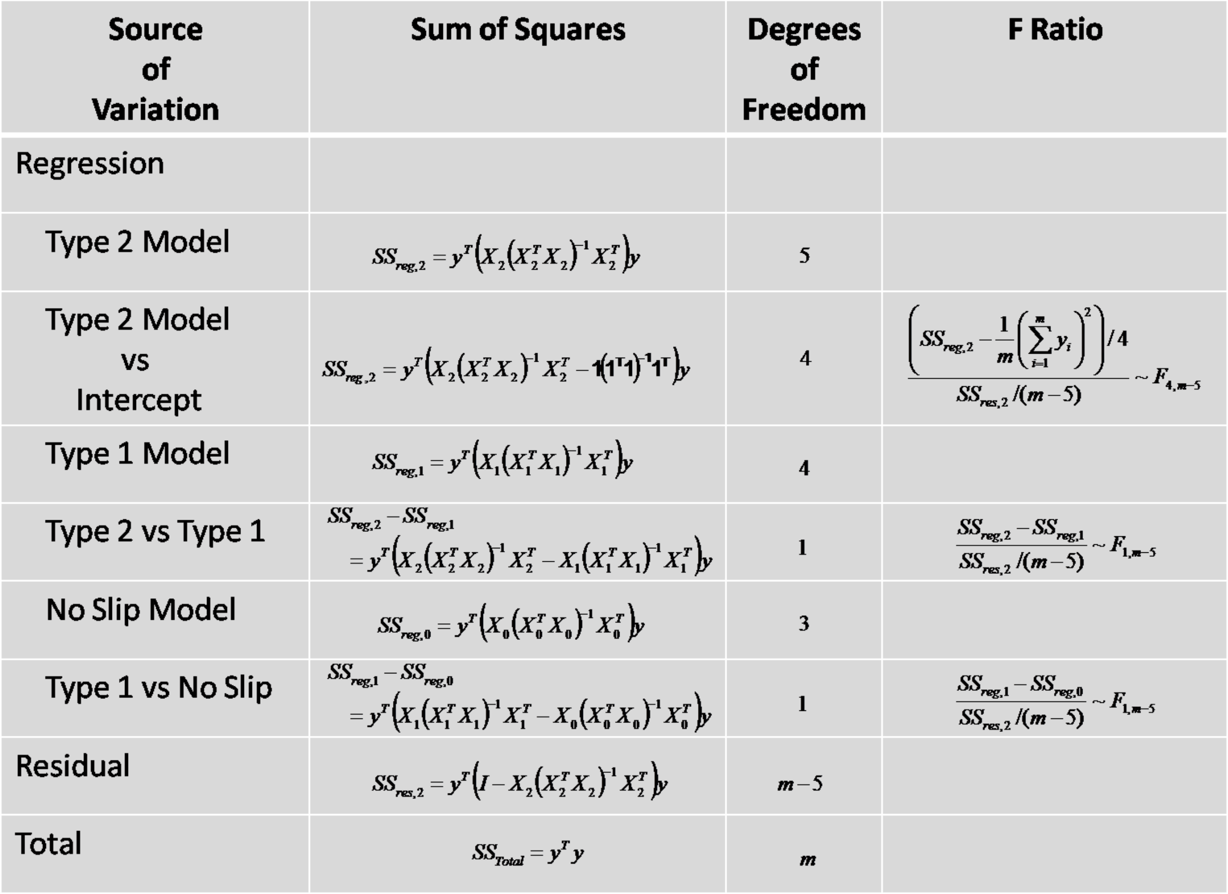
and is the numerator of the statistic in equation 2.3.4-1. As shown in the appendix, this is central under the hypothesis given in equation 2.3.4-2. Thus, under the hypothesis that H_2 is not a correct model for the data, the test statistic is central F-distributed with degrees of freedom 5 and $m-5$

$$\frac{m-5}{5}\cdot\frac{SS_{reg,2} - \frac{1}{m}\left(\sum_{i=1}^{m} y_i\right)^2}{SS_{res,2}} \sim F_{4,m-5} \qquad \text{Eq. 2.3.4-7}$$

So a cycle-slip is only looked for in a given neighborhood when

$$\frac{m-5}{5}\cdot\frac{SS_{reg,2} - \frac{1}{m}\left(\sum_{i=1}^{m} y_i\right)^2}{SS_{res,2}} > F_{4,m-5}(1-\alpha) \qquad \text{Eq. 2.3.4-8}$$

for some significance level α. Now three F-tests have been created so that the correct model can be chosen for a neighborhood. The tests and their corresponding properties are summarized in the following ANOVA table.



Source of Variation	Sum of Squares	Degrees of Freedom	F Ratio
Regression			
Type 2 Model	$SS_{reg,2} = y^T\left(X_2\left(X_2^T X_2\right)^{-1} X_2^T\right)y$	5	
Type 2 Model vs Intercept	$SS_{reg,2} = y^T\left(X_2\left(X_2^T X_2\right)^{-1} X_2^T - \mathbf{1}\left(\mathbf{1}^T\mathbf{1}\right)^{-1}\mathbf{1}^T\right)y$	4	$\dfrac{\left(SS_{reg,2} - \frac{1}{m}\left(\sum_{i=1}^{m} y_i\right)^2\right)/4}{SS_{res,2}/(m-5)} \sim F_{4,m-5}$
Type 1 Model	$SS_{reg,1} = y^T\left(X_1\left(X_1^T X_1\right)^{-1} X_1^T\right)y$	4	
Type 2 vs Type 1	$SS_{reg,2} - SS_{reg,1}$ $= y^T\left(X_2\left(X_2^T X_2\right)^{-1} X_2^T - X_1\left(X_1^T X_1\right)^{-1} X_1^T\right)y$	1	$\dfrac{SS_{reg,2} - SS_{reg,1}}{SS_{res,2}/(m-5)} \sim F_{1,m-5}$
No Slip Model	$SS_{reg,0} = y^T\left(X_0\left(X_0^T X_0\right)^{-1} X_0^T\right)y$	3	
Type 1 vs No Slip	$SS_{reg,1} - SS_{reg,0}$ $= y^T\left(X_1\left(X_1^T X_1\right)^{-1} X_1^T - X_0\left(X_0^T X_0\right)^{-1} X_0^T\right)y$	1	$\dfrac{SS_{reg,1} - SS_{reg,0}}{SS_{res,2}/(m-5)} \sim F_{1,m-5}$
Residual	$SS_{res,2} = y^T\left(I - X_2\left(X_2^T X_2\right)^{-1} X_2^T\right)y$	$m-5$	
Total	$SS_{Total} = y^T y$	m	

Figure 2.3.4-1

It remains to be shown empirically that these tests outperform various real-time slip detection methods, although it is certainly possible that data can appear to behave like a cycle-slip in real-time while the future data indicates that it is the beginning of something else, such as a sudden steep slope or increase in noise.

2.4 Cycle-Slip Repair Tool (CSR)

This section describes how the Cycle-Slip Repair Tool (CSR) implements the methods developed in section 2.3 to detect and repair cycle-slips. Some design decisions are discussed and sample results are presented.

2.4.1 Cycle-Slip Repair Tool Design

The input to CSR is a RINEX file, which contains L1 and L2 signal data for all satellites at a receiver and spans one day. For detailed information on RINEX files see Gurtner & Estey (2007). The CSR consists of three main parts. The first is a RINEX file reader and splitter which separates RINEX files by satellite PRN making one smaller "split RINEX" file for each satellite. This allows the process to be run incrementally or to be run on only a subset of the satellites in the RINEX file. The second part reads a split RINEX file, calculates the iono-free pseudorange (IFPR) and delay from the L1 and L2 signal data in the file, performs cycle-slip repairs on the delay data, recalculates L1 and L2 based on the repaired data, then creates a repaired split RINEX file. The third part of the tool recombines the repaired split RINEX files into a repaired RINEX file for all satellites.

The second part is of interest here. Converting back and forth from L1 and L2 to IFPR and delay is done by applying the transformation in equations 2.2.1-4 and 2.2.1-5. Once the delay data is available, every point must be checked for a cycle-slip. The program implements a moving window of size $2n$. The center of the window is x_0, the point being checked for a slip.

The window itself is the neighborhood $X = \{(x_{-n}, \ldots, x_0, \ldots, x_{n-1}\}$. The window is moved across the delay data set until every point has been tested for a slip using the F-Tests defined in section 2.3. When a cycle-slip is detected, the appropriate model is used to determine the offset(s) and repair the data that follows. Because of the window size, the first and last n points are not checked for slips. For these points, it is assumed that there is not enough local data to determine if there is a cycle-slip.

2.4.2 Sample Results for CSR

The CSR tool was run with a window size of $n = 12$. The threshold is set at an α-level of 0.001 for detecting type 2 slips versus type 1 and 0.0001 for detecting type 1 slips versus no slip. Because of the amount of data being tested, a very small α-level is needed to avoid having many false slip detections. The results below are for the receiver Billings A and the satellite PRN 1.[4] Figure 2.4.2-1 shows the raw delay data for the day of January 1 2000. The satellite is in view of the station at two different times during the day. Only the second track will be examined and is shown in Figure 2.4.2-2.

[4] Billings A is the first of three antennas along with Billings B and Billings C at the WAAS Reference Station (WRS) in Billings, Montana. Each antenna receives the L1 and L2 signals from the satellites in view and so has a delay measurement. Billings is used here because the RINEX file is listed first alphabetically. The GPS satellites are identified by PRN numbers ranging from 1 to 32. PRN 1 is used here because it is the first numerically.

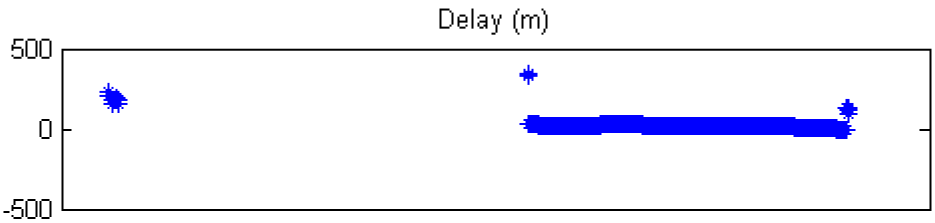


Figure 2.4.2-1

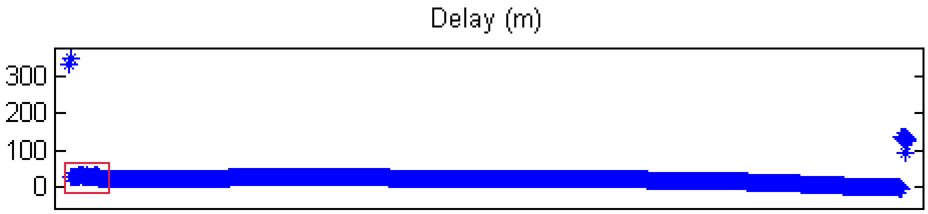


Figure 2.4.2-2

As is normal, there are numerous cycle-slips near the beginning and end of the track when the satellite is at the lowest elevations. The region with the greatest number of slips is highlighted and is shown zoomed below. The repaired delay is also given.

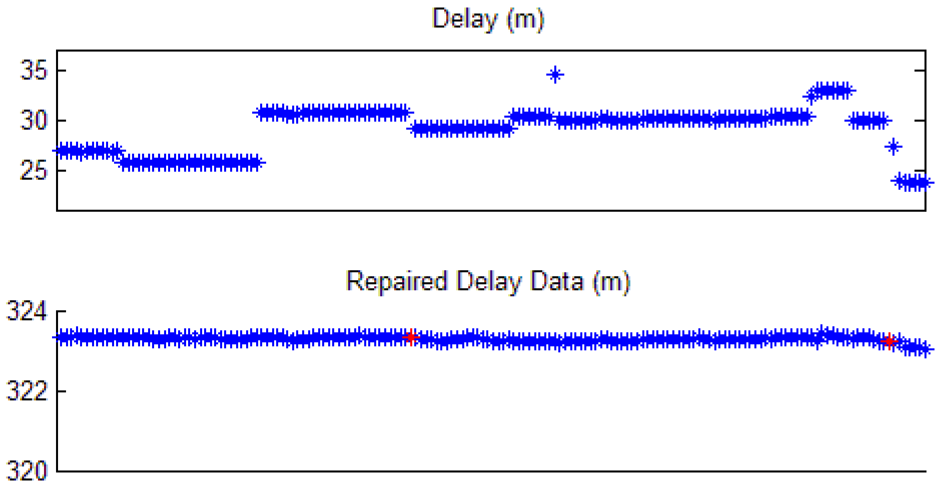


Figure 2.4.2-3

In this portion of the track, there are eight cycle-slips that can be seen in the delay data. Five of these are type 1 cycle-slips and three of them are type 2. In the repaired delay data, the CSR tool has repaired each of the cycle-slips. The repaired values have an added ambiguity because it is not known which portion of the original track, if any, holds correct delay values. So the entire track is arbitrarily brought to match the level that it was at before the first cycle-slip. So the repaired data in Figure 2.4.2-3 does not truly indicate that the delays range between 322 and 324 meters. Remember that the data here is from the carrier phase and that the purpose is ultimately to level the repaired carrier phase onto the code phase, which does not suffer from cycle-slips, as described in section 2.2.2-1. Since the purpose is to level the repaired signal onto the code phase, it is not necessary to resolve the ambiguity that exists in the carrier phase. For carrier phase, only the shape and the continuity achieved by removing cycle-slips are relevant.

In the original data, this portion of the track contains two missing data points. Missing data points are easily handled with this method because there is no requirement that the points in the neighborhood are equally spaced. This holds true even for regions with multiple consecutive missing data points. Additionally, the CSR tool has added these points, plotted in red, by simply interpolating using the linear regression model under H_0.

2.5 Conclusion

A sequential hypothesis test has been devised for detecting and repairing cycle-slips corresponding to various models which are fit using linear regression. This approach allows the addition of any number of cycle-slip models which might occur in ionospheric delay data. The formulated tests yield easy to use results in which the detection of a cycle-slip is dependent on the analysis of variance. The model offset parameter(s) are the only information needed to perform a repair. This method easily handles missing data points and can be used to interpolate values at such locations. Finally, a software tool, the CSR, was created which implements the method to repair RINEX files for use with other analytical tools.

Appendix

Here it is proven that $\frac{2p_{\vec{e},\vec{n}}\,p_{\vec{n},\vec{u}}\,p_{\vec{e},\vec{u}} - p_{\vec{n},\vec{u}}^2 - p_{\vec{e},\vec{u}}^2}{1 - p_{\vec{e},\vec{n}}^2} \le 0$ as stated in section 1.4.3. This shows that VDOP is bounded as claimed. The other DOPs have similar terms with the $\vec{e}$, $\vec{n}$, and $\vec{u}$ vectors permutated. It is assumed that $\left|\rho_{\vec{e},\vec{n}}\right| \ne 1$. For this assumption to fail, the geometry matrix would have to be singular.

Lemma A-1: $\frac{2p_{\vec{e},\vec{n}}\,p_{\vec{n},\vec{u}}\,p_{\vec{e},\vec{u}} - p_{\vec{n},\vec{u}}^2 - p_{\vec{e},\vec{u}}^2}{1 - p_{\vec{e},\vec{n}}^2} \le 0$ for any $\vec{e}$, $\vec{n}$, and $\vec{u}$ vectors.

pf: This requires only a simple algebraic manipulation while recalling that $-1 \le \rho_{\vec{x},\vec{y}} \le 1$ for any vectors $\vec{x}$ and $\vec{y}$.

$$\begin{aligned}
\frac{2p_{\vec{e},\vec{n}}\,p_{\vec{n},\vec{u}}\,p_{\vec{e},\vec{u}} - p_{\vec{n},\vec{u}}^2 - p_{\vec{e},\vec{u}}^2}{1 - p_{\vec{e},\vec{n}}^2} &= \frac{2p_{\vec{e},\vec{n}}\,p_{\vec{n},\vec{u}}\,p_{\vec{e},\vec{u}} - \left|p_{\vec{n},\vec{u}}\right|^2 - \left|p_{\vec{e},\vec{u}}\right|^2}{1 - \left|p_{\vec{e},\vec{n}}\right|^2} \\
&\le \frac{2 \cdot \left|p_{\vec{e},\vec{n}}\right| \cdot \left|p_{\vec{n},\vec{u}}\right| \cdot \left|p_{\vec{e},\vec{u}}\right| - \left|p_{\vec{n},\vec{u}}\right|^2 - \left|p_{\vec{e},\vec{u}}\right|^2}{1 - \left|p_{\vec{e},\vec{n}}\right|^2} \\
&\le \frac{2 \cdot \left|p_{\vec{n},\vec{u}}\right| \cdot \left|p_{\vec{e},\vec{u}}\right| - \left|p_{\vec{n},\vec{u}}\right|^2 - \left|p_{\vec{e},\vec{u}}\right|^2}{1 - \left|p_{\vec{e},\vec{n}}\right|^2} \\
&= \frac{-\left(\left|p_{\vec{n},\vec{u}}\right| - \left|p_{\vec{e},\vec{u}}\right|\right)^2}{1 - \left|p_{\vec{e},\vec{n}}\right|^2} \\
&\le 0
\end{aligned}$$

Q.E.D.

Now it is proven that the matrices used in section 2.3.3 called A_1, A_2, A_3, and A_4 are symmetric and idempotent as claimed.

Lemma A-2: Any matrix of the form $A = B(B^T B)^{-1} B^T$ is symmetric and idempotent.

pf: Symmetry can be proven by simply showing that $A^T = A$.

$$A^T = \left(B(B^T B)^{-1} B^T\right)^T = \left(B^T\right)^T \left((B^T B)^{-1}\right)^T B^T = B\left(\left(B^T B\right)^T\right)^{-1} B^T = B\left(B^T B\right)^{-1} B^T = A$$

Idempotency can be proven by simply showing that $A^2 = A$ so that A is idempotent by definition.

$$A^2 = \left(B(B^T B)^{-1} B^T\right)^2 = B(B^T B)^{-1} B^T B(B^T B)^{-1} B^T = B(B^T B)^{-1} B^T = A$$

Q.E.D.

Lemma A-3: Any matrix of the form $A = I - B(B^T B)^{-1} B^T$ is symmetric and idempotent.

pf: Symmetry follows directly from $B(B^T B)^{-1} B^T$ and I being symmetric. Idempotency is again demonstrated by showing that $A^2 = A$.

$$\begin{aligned} A^2 &= \left(I - B(B^T B)^{-1} B^T\right)^2 \\ &= I - 2B(B^T B)^{-1} B^T + B(B^T B)^{-1} B^T B(B^T B)^{-1} B^T \\ &= I - 2B(B^T B)^{-1} B^T + B(B^T B)^{-1} B^T \\ &= A \end{aligned}$$

Q.E.D.

Lemma A-4: Any matrix of the form $A = B(B^T B)^{-1} B^T - C(C^T C)^{-1} C^T$ is symmetric.

pf: This follows directly from the symmetry of $B(B^T B)^{-1} B^{T^T}$ and $C(C^T C)^{-1} C^T$ proven above in claim 1.

Lemma A-5: Any matrix of the form $A = B(B^T B)^{-1} B^T - C(C^T C)^{-1} C^T$ is idempotent if $B = [C \mid S]$ for some matrix S.

pf: Let $B = [C \mid S]$ and consider the partitioned form:

$$\begin{aligned}
\left[\frac{C^T B(B^T B)^{-1} B^T}{S^T B(B^T B)^{-1} B^T}\right] &= \left[\frac{C^T}{S^T}\right] + \left[\frac{C^T\left(B(B^T B)^{-1} B^T - I\right)}{S^T\left(B(B^T B)^{-1} B^T - I\right)}\right] \\
&= \left[\frac{C^T}{S^T}\right] + \left[\frac{C^T}{S^T}\right]\left(B(B^T B)^{-1} B^T - I\right) \\
&= \left[\frac{C^T}{S^T}\right] + B^T\left(B(B^T B)^{-1} B^T - I\right) \\
&= \left[\frac{C^T}{S^T}\right]
\end{aligned}$$

From the top partitions on the left and right side this implies that

$$C^T B(B^T B)^{-1} B^T = C^T \qquad \text{Eq. A5-1}$$

or from transposing that

$$B(B^T B)^{-1} B^T C = C \qquad \text{Eq. A5-2}$$

Now, squaring $A^2 = B(B^T B)^{-1} B^T - C(C^T C)^{-1} C^T$ yields

$$\begin{aligned} A^2 &= \left(B(B^T B)^{-1} B^T - C(C^T C)^{-1} C^T\right)^2 \\ &= \left(B(B^T B)^{-1} B^T\right)^2 - B(B^T B)^{-1} B^T C(C^T C)^{-1} C^T \\ &\quad - C(C^T C)^{-1} C^T B(B^T B)^{-1} B^T + \left(C(C^T C)^{-1} C^T\right)^2 \end{aligned}$$

Substituting equations A5-1 and A5-2 into the two middle terms gives

$$A^2 = \left(B(B^T B)^{-1} B^T\right)^2 - C(C^T C)^{-1} C^T - C(C^T C)^{-1} C^T + \left(C(C^T C)^{-1} C^T\right)^2$$

So, from claim 1

$$\begin{aligned} A^2 &= B(B^T B)^{-1} B^T - C(C^T C)^{-1} C^T - C(C^T C)^{-1} C^T + C(C^T C)^{-1} C^T \\ &= B(B^T B)^{-1} B^T - C(C^T C)^{-1} C^T \\ &= A \end{aligned}$$

Q.E.D.

Finally, it is shown that the quadratic forms Q_2, Q_3 and Q_4 have non-centrality parameters $\lambda_2 = \lambda_3 = \lambda_4 = 0$ as claimed in section 2.3.3.

Lemma A-6: The quadratic form Q_2 has non-centrality parameter $\lambda_2 = 0$.

pf: According to the Cochran-Fisher theorem, the non-centrality parameter of Q_2 is

$\lambda_2 = \frac{1}{2}\vec{\mu}^T A_2 \vec{\mu}$ where $\mu = E\left(\frac{\vec{y}}{\sigma}\right) = \frac{X_2\beta}{\sigma}$. So,

$$\begin{aligned}
\lambda_2 &= \frac{1}{2\sigma^2}(X_2\beta)^T\left(I - X_2(X_2^T X_2)^{-1} X_2^T\right)(X_2\beta) \\
&= \frac{1}{2\sigma^2}\left(\beta^T X_2^T X_2\beta - \beta^T X_2^T X_2 (X_2^T X_2)^{-1} X_2^T X_2\beta\right) \\
&= \frac{1}{2\sigma^2}\left(\beta^T X_2^T X_2\beta - \beta^T X_2^T X_2\beta\right) \\
&= 0
\end{aligned}$$

Q.E.D.

Lemma A-7: Let $Q = \left(\frac{\vec{y}}{\sigma}\right)^T \left(A(A^T A)^{-1} A^T - B(B^T B)^{-1} B^T\right)\left(\frac{\vec{y}}{\sigma}\right)$ be a chi-squared quadratic form where A is $m \times p$, B is $m \times p - k$, and . $A = [B \mid C]$ for some matrix C that is $m \times k$. Let $\vec{y} \sim N(A\beta, \sigma^2)$ and $\vec{\beta} = \begin{bmatrix} \vec{\beta}_1 \\ \hline \vec{\beta}_2 \end{bmatrix}$ where $\vec{\beta}_1$ is $p - k \times 1$ and $\vec{\beta}_2$ is $k \times 1$. Q has non-centrality parameter $\lambda = 0$ under the hypothesis $\vec{\beta}_2 = \mathbf{0}_\mathrm{k}$.

pf: The non-centrality parameter of Q is

$$\lambda = \frac{1}{2}\vec{\mu}^T\left(A(A^T A)^{-1} A^T - B(B^T B)^{-1} B^T\right)\vec{\mu}$$

where $\vec{\mu} = E\left(\frac{\vec{y}}{\sigma}\right) = \frac{A\beta}{\sigma}$. So, noting that

$$A\beta = \left[B \mid C\right]\begin{bmatrix}\vec{\beta}_1 \\ \vec{\beta}_2\end{bmatrix} = \left[B\vec{\beta}_1 + C\vec{\beta}_2\right]$$

yields

$$\begin{aligned}
\lambda &= \frac{1}{2\sigma^2}\left(A\vec{\beta}\right)^T\left(A(A^TA)^{-1}A^T - B(B^TB)^{-1}B^T\right)\left(A\vec{\beta}\right) \\
&= \frac{1}{2\sigma^2}\vec{\beta}^T A^T\left(A(A^TA)^{-1}A^T - B(B^TB)^{-1}B^T\right)A\vec{\beta} \\
&= \frac{1}{2\sigma^2}\left(\vec{\beta}^T A^T A(A^TA)^{-1}A^T A\vec{\beta} - \vec{\beta}^T A^T B(B^TB)^{-1}B^T A\vec{\beta}\right) \\
&= \frac{1}{2\sigma^2}\left(\vec{\beta}^T A^T A\vec{\beta} - \vec{\beta}^T A^T B(B^TB)^{-1}B^T A\vec{\beta}\right) \\
&= \frac{1}{2\sigma^2}\left(\left[B\vec{\beta}_1 + C\vec{\beta}_2\right]^T\left[B\vec{\beta}_1 + C\vec{\beta}_2\right] - \left[B\vec{\beta}_1 + C\vec{\beta}_2\right]^T B(B^TB)^{-1}B^T\left[B\vec{\beta}_1 + C\vec{\beta}_2\right]\right) \\
&= \frac{1}{2\sigma^2}\left(\left[\vec{\beta}_1^T B^T + \vec{\beta}_2^T C^T\right]\left[B\vec{\beta}_1 + C\vec{\beta}_2\right] - \left[\vec{\beta}_1^T B^T + \vec{\beta}_2^T C^T\right]B(B^TB)^{-1}B^T\left[B\vec{\beta}_1 + C\vec{\beta}_2\right]\right) \\
&= \frac{1}{2\sigma^2}\left(\left[\vec{\beta}_1^T B^T + \vec{\beta}_2^T C^T\right]\left[B\vec{\beta}_1 + C\vec{\beta}_2\right] - \left[\vec{\beta}_1^T B^T B(B^TB)^{-1}B^T + \vec{\beta}_2^T C^T B(B^TB)^{-1}B^T\right]\left[B\vec{\beta}_1 + C\vec{\beta}_2\right]\right) \\
&= \frac{1}{2\sigma^2}\left(\left[\vec{\beta}_1^T B^T + \vec{\beta}_2^T C^T\right]\left[B\vec{\beta}_1 + C\vec{\beta}_2\right] - \left[\vec{\beta}_1^T B^T + \vec{\beta}_2^T C^T B(B^TB)^{-1}B^T\right]\left[B\vec{\beta}_1 + C\vec{\beta}_2\right]\right)
\end{aligned}$$

Under the hypothesis that $\vec{\beta}_2 = \mathbf{0}_{\mathrm{k}}$, this gives:

$$\begin{aligned}
\lambda &= \frac{1}{2\sigma^2}\left(\left[\vec{\beta}_1^T B^T + \mathbf{0}_k^T C^T\right]\left[B\vec{\beta}_1 + C\mathbf{0}_k\right] - \left[\vec{\beta}_1^T B^T + \mathbf{0}_k^T C^T B(B^TB)^{-1}B^T\right]\left[B\vec{\beta}_1 + C\mathbf{0}_k\right]\right) \\
&= \frac{1}{2\sigma^2}\left(\vec{\beta}_1^T B^T B\vec{\beta}_1 - \vec{\beta}_1^T B^T B\vec{\beta}_1\right) \\
&= 0
\end{aligned}$$

Q.E.D.

Reference List

Banville, S., & Langley, R. B. (2010). Instantaneous cycle-slip correction for real-time PPP applications. *Navigation: Journal of the Institute of Navigation, Vol. 57,* No. 4, 325-334.

Ben-Israel, A. (1992). A volume associated with m x n matrices. *Linear Algebra and its Applications, 167*, 87-111.

Chan, N. N. (1982). A-Optimality for regression designs. *Journal of Mathematical Analysis and Applications*, *87*, 45-50.

Chan, N. N., & Wong, C. S. (1980). *Existence of an A-optimal model for the regression experiment. Journal of Mathematical Analysis and Applications*, *77*, 403-415.

Fallat, S. M., & Johnson, C. R. (2000). Determinantal inequalities: Ancient history and recent advances. *Contemporary Mathematics*, *Vol. 250*, 199-212.

Farrell, J. A. (2008). *Aided navigation: GPS with high rate sensors*. McGraw-Hill.

Farrell, J. A., & Barth, M. (1998). *The Global Positioning System & inertial navigation*. McGraw-Hill.

Gover, E., & Krikorian, N. (2010). Determinants and the volume of parallelotopes and zonotopes. *Linear Algebra and its Applications*, doi:10.1016/j.laa.2010.01.031

Deza, M., & Grishukhin, V. (2004). Properties of parallelotopes equivalent to Voronoi's conjecture. *European Journal of Combinatronics*, 25, 518-533. doi:10.1016/j.ejc.2003.09.012

Gurtner, W., Estey, L. (2007). RINEX the receiver independent exchange format version 3.00. Astronomical Institute University of Bern and UNAVCO. Retrieved from ftp://ftp.unibe.ch/aiub/rinex/rinex300.pdf.

Hooks, T., Marx, D., Kachman, S., & Pederson, J. (2009). Optimality criteria for models with random effects. *Revista Colombiana de Estadistica*, *32*, no. 1, 17-31.

Jones, F. (2011). *Math 221/222 Honors Calculus III/IV*. Rice University. Retrieved from http://math.rice.edu/Courses/homepages/Math221-222.html.

Kaplan, E. D., & Hegarty, C. J. (2006). *Understanding GPS principles and applications* (2nd ed.). Norwood, Massachusetts: Artech House, Inc.

Langley, R. B. (1999). Dilution of Precision. *GPS World*, *May 1999*, 52-59.

Leng, G., & Zhang, Y. (1998). The generalized sine theorem and inequalities for simplices. *Linear Algebra and its Applications*, *278*, 237-247.

Mackay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge, United Kingdom: Cambridge University Press.

Marquis, C. W. (1993) Integration of differential GPS and inertial navigation using a complimentary Kalman filter. (Master's thesis). Naval Postgraduate School, Monterey, California. Retrieved from Defense Technical Information Center.

Marujao, & D. B., Mendes, V. B. (2007, September). Investigation of instantaneous ambiguity resolution with GPS / Galileo combination for precise kinematic positioning. Paper presented at ION GNSS 20th International Technical Meeting of the Satellite Division, Fort Worth, TX.

Miao, J., & Ben-Isreal, A. (1991). On principal angles between subspaces in $\Re^n$. *Linear Algebra and its Applications, 171*, 81-98.

Misra, P., & Enge, P. (2006). *Global Positioning System signals, measurements, and performance* (2nd ed.). Lincoln, Massachusetts: Ganga-Jamuna Press.

Myers, R. H., & Milton, J. S. (1998). A *first course in the theory of linear statistical models*. The McGraw-Hill Companies, Inc.

Nakos, G., Joyner, D. (1998). *Linear algebra with applications*. Pacific Grove, California: Brooks/Cole Publishing Company. pp. 402-403.

Parkinson, B. W., & Spilker, J. J. (1996). *Global Positioning System: Theory and applications volume 1*. American Institute of Aeronautics and Astronautics, Inc.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). John Wiley & Sons.

Schnell, U. (1993). Volumes of projections of parallelotopes. *Bull. London Math. Soc.*, *26* (1994), 181-184.

Teunissen, P. J. G. (2003a). An invariant upperbound for the GNSS bootstrappend ambiguity success rate. *Journal of Global Positioning Systems, Vol. 2*, No. 1, 13-17.

Teunissen, P. J. G. (2003b). *Towards a unified theory of GNSS ambiguity resolution. Journal of Global Positioning Systems, Vol. 2*, No. 1, 1-12.