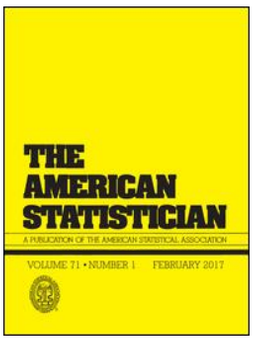


Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

# An Intuitive Geometric Approach to the Gauss Markov Theorem

Leandro da Silva Pereira, Lucas Monteiro Chaves & Devanil Jaques de Souza

**To cite this article:** Leandro da Silva Pereira, Lucas Monteiro Chaves & Devanil Jaques de Souza (2017) An Intuitive Geometric Approach to the Gauss Markov Theorem, The American Statistician, 71:1, 67-70, DOI: 10.1080/00031305.2016.1209127

**To link to this article:** http://dx.doi.org/10.1080/00031305.2016.1209127

Accepted author version posted online: 21 Jul 2016.
Published online: 21 Jul 2016.

THE AMERICAN STATISTICIAN
2017, VOL. 71, NO. 1, 67–70
http://dx.doi.org/10.1080/00031305.2016.1209127

Taylor & Francis
Taylor & Francis Group

# An Intuitive Geometric Approach to the Gauss Markov Theorem

Leandro da Silva Pereira[a], Lucas Monteiro Chaves[b], and Devanil Jaques de Souza[b]

[a]UTFPR – Federal Technological University of Parana, DAMAT, Apucarana-PR, Brazil; [b]UFLA – Federal University of Lavras, DEX, Lavras-MG, Brazil

**ABSTRACT**

Algebraic proofs of Gauss–Markov theorem are very disappointing from an intuitive point of view. An alternative is to use geometry that emphasizes the essential statistical ideas behind the result. This article presents a truly geometrical intuitive approach to the theorem, based only in simple geometrical concepts, like linear subspaces and orthogonal projections.

**ARTICLE HISTORY**
Received November 2014
Revised June 2016

**KEYWORDS**
Dispersion cloud of points;
Gauss–Markov estimator;
Orthogonal projection



## 1. Introduction

There are few general results in statistics. Often, particular and very restrictive assumptions are necessary, like, for example, normality. One of these few general results is the Gauss–Markov theorem, with highly practical consequences. The majesty of Gauss–Markov theorem relies in the fact that it holds regardless the distribution of the random variable considered. This result gives statistical meaning to a pure mathematical fact: the least-square method. Since the theorem is a basic result, it is taught in beginning statistics courses. However, the proofs presented in most of the textbooks (Rao 1999; Casella and Berger 2002; Rencher 2008) are based only on algebraic properties of positive definite matrices. The experience in class seems to lead us to conclude that this kind of demonstration does not improve student comprehension of the result. Proving the Gauss–Markov theorem by algebraic methods seems to be at most innocuous. There are demonstrations that adopt a geometric flavor, but they usually rely on some algebraic results (Saville and Wood 1991; Gruber 1998; Ruud 2000). First of all, we have to be clear about what is meant by a geometric demonstration. In general, this will not be, from a mathematical point of view, a totally rigorous proof, requiring some degree of intuition. The tricky part of our geometric approach is to interpret matrices as linear transformations. Linear transformations can be viewed as geometric objects because, as functions, they transform vectors in vectors and linear subspaces in linear subspaces. In that sense it is a very concrete geometrical object. Another basic geometric concept is the vector projection onto linear subspaces, particularly the orthogonal projections. The purpose of this article is to provide an intuitive geometric demonstration of the Gauss–Markov theorem, using only the concepts of linear subspaces, linear transformations, and projections.

## 2. The Linear Model

Consider $\mathbf{Y}$ a random vector with unknown mean vector $\boldsymbol{\mu} = E[\mathbf{Y}]$. By reasons related to the random experiment that generate $\mathbf{Y}$, we can suppose some linear relations among the components of the unknown mean vector $\boldsymbol{\mu}$ and, therefore, assume that the vector $\boldsymbol{\mu}$ belongs to some known linear subspace $\mathbf{W}$, which, in essence, characterizes a linear model. The vector $\mathbf{Y}$ stands in the data space, in general, the $n$-dimensional Euclidean space $\mathbb{R}^n$. Since the dimension of data space is higher than the dimension of $\mathbf{W}$, that is, there are more data than characteristics to be estimated, it is plausible to use a lower number of variables to describe the $\mathbf{W}$ subspace. Such procedure is called parameterization and can be done in the following way: consider $\mathbf{W}$ to be the image of a linear transformation $\mathbf{X}$, defined in another vector space, which will be called parameter space, $\mathbf{W} = \mathrm{Im}(\mathbf{X})$. To avoid technical difficulties, the linear transformation $\mathbf{X}$ will be considered injective, that is, for each vector $\mathbf{w}$ in $\mathbf{W}$, there exists a unique vector $\boldsymbol{\beta}$ in the parameter space such that $\mathbf{w} = \mathbf{X}\boldsymbol{\beta}$. In practical situations, the experiment defines the matrix $\mathbf{X}$, the design matrix, and the column subspace of $\mathbf{X}$ defines $\mathbf{W}$. Then it is possible to make clear the linear model assumptions: $\mathbf{Y}$ is a random vector in the data space, $\boldsymbol{\mu} = \mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ a vector in $\mathbf{W}$, where $\boldsymbol{\beta}$ is an unknown vector in the parameter space and $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is the vector of errors.

All of this can be described geometrically by Figure 1.

The greatest advantage of representing the linear model as in Figure 1 is the description of the estimation process. What is the estimation process? It is a very simple decision: after a data vector $\mathbf{y}$ is observed, we have only to choose a vector in $\mathbf{W}$, which we believe to be a good representative of $\mathrm{E}[\mathbf{Y}]$. If $\mathbf{y}$ belongs to the space $\mathbf{W}$, as this space is the space where the mean of the random vector $\mathbf{Y}$ is restricted, there is no reason to not estimate $\mathrm{E}[\mathbf{Y}]$ by the observed vector $\mathbf{y}$. But this seldom occurs, since the observed vector $\mathbf{y}$ is affected by random errors. Therefore, almost surely, the vector $\mathbf{y}$ does not belong to the subspace $\mathbf{W}$ and a natural procedure to estimate the vector $\mathrm{E}[\mathbf{Y}]$ is to take some kind of projection of $\mathbf{y}$ on $\mathbf{W}$. That process, when some specific restrictions on the projection are considered, explains more sophisticated estimation methods, like ridge regression or elastic net regression (Hoerl and Kennard 1970; Zou and Hastie 2005). Let us go to the simplest idea: to choose the vector in $\mathbf{W}$

**CONTACT** Leandro da Silva Pereira *leandropereira@utfpr.edu.br* UTFPR – Federal Technological University of Parana, DAMAT, Apucarana-PR 86812-460, Brazil.

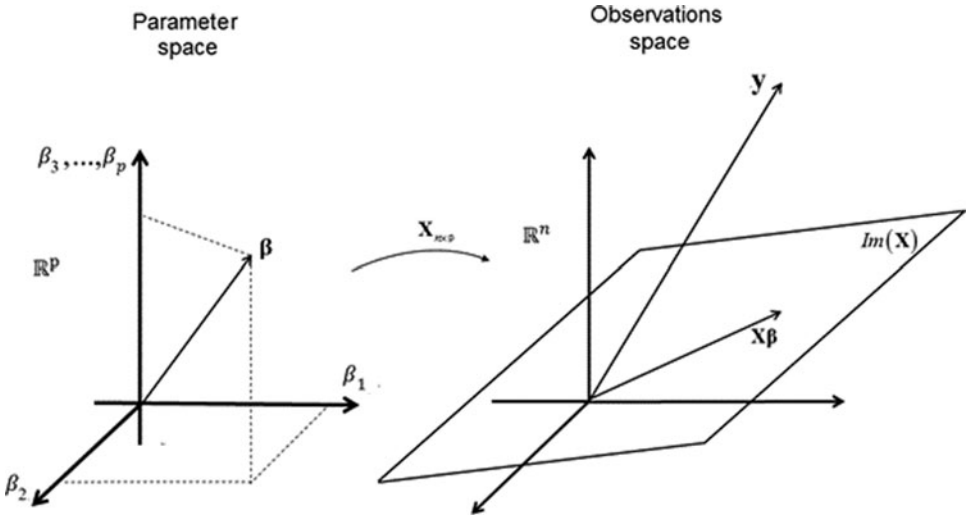


**Figure 1.** Geometrical characterization of a linear model.



closest to $\mathbf{y}$. This estimation procedure is called *the least-square method*. How to do this? The answer is to use linear orthogonal projections. If $P_W$ is the linear orthogonal projection onto $\mathbf{W}$, the chosen vector is $P_W(\mathbf{y})$ (Rencher 2008, p. 43, p. 228). Since the linear transformation $\mathbf{X}$ is injective, there is only one $\hat{\boldsymbol{\beta}}$ such that $P_W(\mathbf{y}) = \mathbf{X}\hat{\boldsymbol{\beta}}$. This equation, in its algebraic form, is denominated *the normal equation* and $\hat{\boldsymbol{\beta}}$ is the least-square estimate of the parameter vector $\boldsymbol{\beta}$. To express $\hat{\boldsymbol{\beta}}$ in terms of the data vector $\mathbf{y}$, it is necessary to have an expression of the projection $P_W$ as a matrix. This can be done with a little linear algebra, getting $P_W = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, where $\mathbf{X}'$ is the transpose of $\mathbf{X}$. But that linear algebra is not necessary to understand what follows.

## 3. The Geometry of Gauss–Markov Theorem

The statistical properties of such estimation method are established by the Gauss–Markov theorem. Recall that, given a random vector $\mathbf{Y}$ with covariance matrix $D(\mathbf{Y}) = E[(\mathbf{Y} - E[\mathbf{Y}])(\mathbf{Y} - E[\mathbf{Y}])']$, the total variance is the sum of the variances of each component of $\mathbf{Y}$, that is, the trace of $D(\mathbf{Y})$.

*Theorem 1 (Gauss–Markov).* If $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with covariance matrix $D(\mathbf{Y}) = \sigma^2\mathbf{I}$, then the least-square estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ has minimum total variance among all linear unbiased estimators of $\boldsymbol{\beta}$.

First of all, it is necessary to have an intuitive idea of the meaning of "covariance matrix." If a lot of values of the random vector $\mathbf{Y}$ are observed, they form a cloud of points in the data space. We will call this cloud the dispersion cloud. The matrix $D(\mathbf{Y})$ tells us about the shape of this cloud. This can be seen in the following way. Consider a unitary vector $\mathbf{x}$. The orthogonal projection of the random vector $\mathbf{Y}$ in the direction of $\mathbf{x}$ defines a one-dimensional random variable given by $\mathbf{Y} \cdot \mathbf{x} = \|\mathbf{Y}\| \cos(\theta)$, where $\mathbf{Y} \cdot \mathbf{x}$ is the inner product and $\theta$ is the angle between $\mathbf{x}$ and $\mathbf{Y}$. In this way, we have a one-dimensional random variable with the same direction as $\mathbf{x}$ and centered on $\boldsymbol{\mu} \cdot \mathbf{x} = \|\boldsymbol{\mu}\| \cos(\theta')$, where $\theta'$ is the angle between $\boldsymbol{\mu}$ and $\mathbf{x}$ (Figure 2).

The variance of this random variable gives a good idea of the width of the dispersion cloud in the $\mathbf{x}$ direction. As we are supposing $D(\mathbf{Y}) = \sigma^2\mathbf{I}$, the variance is $\text{var}(\mathbf{x} \cdot \mathbf{Y}) = \mathbf{x}'D(\mathbf{Y})\mathbf{x} = \sigma^2\mathbf{x}'\mathbf{x} = \sigma^2$. That is, the variability does not depend on the direction of $\mathbf{x}$. So, all directions in the data space are equally likely. This means that the width of the dispersion cloud must be almost the same in any direction, that is, the dispersion cloud has approximate spherical symmetry. To delimit the cloud, we can take a sphere that contains, say, approximately 75% of the cloud points. Furthermore, with high probability, such cloud should be closely centered at $\boldsymbol{\mu}$. Observe that, since $\mu \in \mathbf{W}$, this spherical cloud must be almost symmetric in relation to $\mathbf{W}$. If the observed data vectors in the cloud are orthogonally projected into the subspace $\mathbf{W}$, such projection will have an image with almost spherical symmetry and the same radius of the original dispersion cloud.

The task now is to visualize the dispersion cloud of the estimator $\hat{\boldsymbol{\beta}}$. The dispersion cloud of the least-square estimator $\hat{\beta}$ is then obtained by taking, in the parameter space, the preimage, by the transformation $\mathbf{X}$, of the projected dispersion cloud onto $\mathbf{W}$. Since, in general, the transformation $\mathbf{X}$ does not preserve distance, the preimage will not have spherical symmetry anymore. With basic linear algebra knowledge, it is possible to prove that this preimage will have an elliptical symmetry. Then, the dispersion cloud of the least-square estimator $\hat{\boldsymbol{\beta}}$ is, with high probability, approximately an ellipsoid centered in

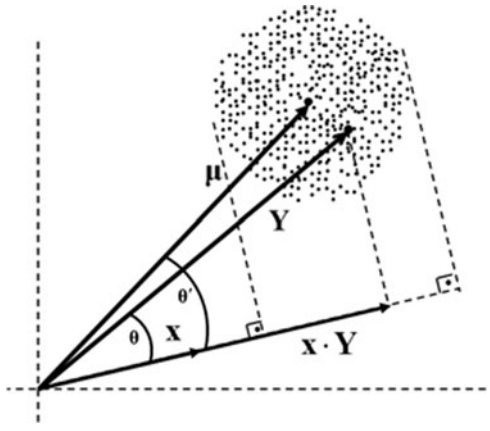


**Figure 2.** Random variable in the *x* vector direction.

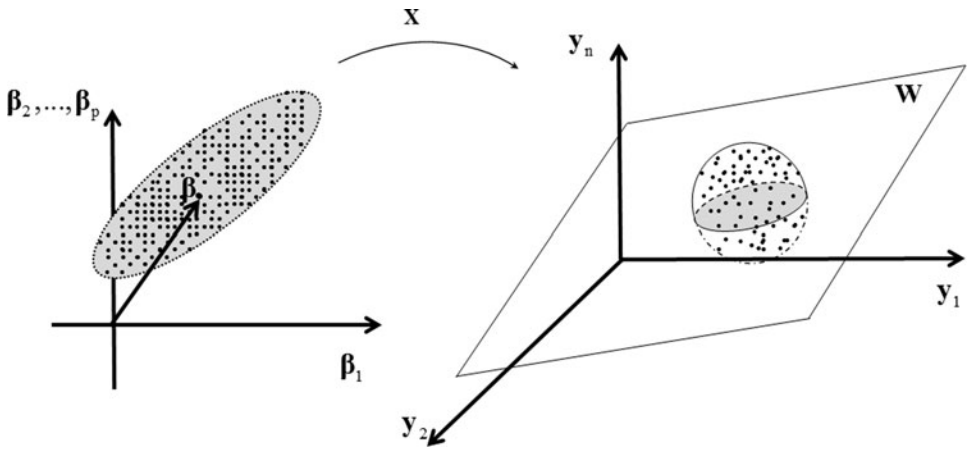


**Figure 3.** Dispersion clouds of the random vectors **Y** and $\hat{\boldsymbol{\beta}}$.

the real vector $\boldsymbol{\beta}$, as shown in Figure 3. In other words, a sphere in **W**, $(\mathbf{w}-\boldsymbol{\mu})'(\mathbf{w}-\boldsymbol{\mu}) = \text{const.}$, is the image, by the transformation **X**, of an ellipse centered at $\boldsymbol{\beta}$, $(\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \text{const.}$

The symmetry of the dispersion cloud of $\hat{\beta}$ gives an intuitive notion that the least-square estimator $\hat{\beta}$ is unbiased, as in fact it is. Students should take away from this development that the concepts of unbiasedness and geometrical symmetry are closely related.

Now let us analyze the behavior of other linear unbiased estimators. Such estimators have the form $\widetilde{\beta} = \mathbf{L}\,\mathbf{Y}$, where **L** is a $p \times n$ matrix and

$$\boldsymbol{\beta} = \mathrm{E}\left[\widetilde{\boldsymbol{\beta}}\right] = \mathrm{E}\left[\mathbf{LY}\right] = \mathbf{L}\,\mathrm{E}\left[\mathbf{Y}\right] = \mathbf{LX}\boldsymbol{\beta}.$$

Since $\boldsymbol{\beta}$ is unknown, the equality must hold for every $\boldsymbol{\beta}$, then, $\mathbf{L}\,\mathbf{X} = \mathbf{I}$. It follows from this equality that $(\mathbf{XL})^2 = (\mathbf{XL})(\mathbf{XL}) = \mathbf{X}(\mathbf{LX})\mathbf{L} = \mathbf{X}(\mathbf{I})\mathbf{L} = \mathbf{XL}$. This property implies that **XL** is a projection onto **W** (Gentle 2007, p. 286). So, any other linear unbiased estimator is obtained in the same way as the least-square estimator, that is, it is obtained as a linear projection onto **W**. The distinction between them is that the projection is no longer orthogonal as in the least-square case. A nonorthogonal projection is denominated *oblique*. We will not give its precise mathematical definition, but it is possible to have a good idea of how it works by only looking at the two-dimensional case (Figure 4).

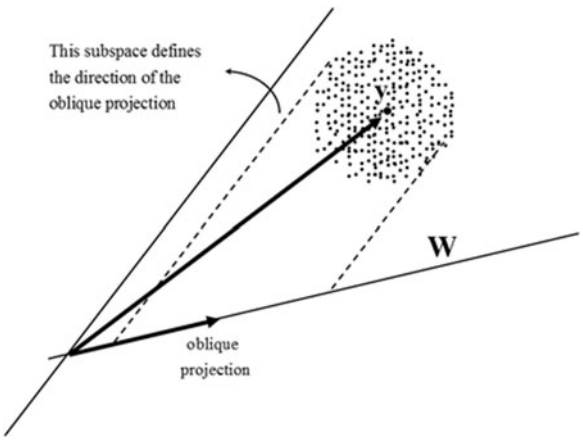


**Figure 4.** Oblique projection.

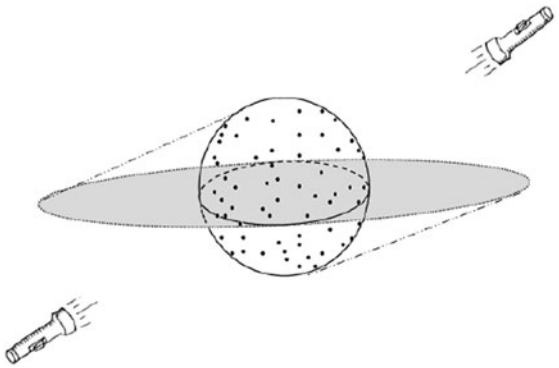

**Figure 5.** Oblique projection of the dispersion cloud.

Remembering once more that the dispersion cloud of **Y** is a sphere approximately centered at the vector $\mathbf{X}\boldsymbol{\beta}$ in **W**, the oblique projection of this sphere is no longer spherical but elliptical. The following analogy helps to see this. The shadow of a sphere, projected by the sun at noon, is circular shaped with unchanged radius. As the sun sets, the projected image becomes an ellipse with increasing eccentricity. Another good idea is to perform an experiment using a flashlight to make the projection and a styrofoam sphere to represent the dispersion cloud (Figure 5).

To compare the total variance of the least-square estimator with the total variance of other linear unbiased estimators, it is enough heuristically to compare the dispersion clouds obtained by orthogonal and oblique projections in the **W** subspace. The spherical cloud related to least-square estimator is entirely contained in the elliptic cloud obtained by oblique projection. Taking the preimage, by **X**, the same occurs for the dispersion cloud in the parameter space. This demonstrates, intuitively, that the total variance of the least-square estimator is lower than the total variance of any other linear unbiased estimator. Therefore, we have an intuitive geometric demonstration of the Gauss–Markov Theorem.

## 4. Conclusions

Although it is known but not often used, geometry is the natural context for problems related to least-square methods. The use of geometrical arguments is intuitive and enlightens the statistical

concepts. We believe that the geometrical approach to visualizing the Gauss–Markov theorem has considerable pedagogical value.

## References

Casella, G., and Berger, R. L. (2002), *Statistical Inference* (2nd ed.), Pacific Grove, CA: Duxbury. [67]

Gentle, E. J. (2007), *Matrix Algebra. Theory, Computations, and Applications in Statistics*, New York: Springer. [69]

Gruber, M. H. J. (1998), *Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators*, New York: Marcel Dekker. [67]

Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression. Biased Estimation for Non-Orthogonal Problems," *Technometrics*, 12, 55–67. [67]

Rao, C. R., and Toutenburg, H. (1999), *Linear Models, Least Squares and Alternatives* (2nd ed.), New York: Springer-Verlag. [67]

Rencher, A. C., and Schaalje, G. B. (2008), *Linear Models in Statistics*, Hoboken, NJ: Wiley. [67]

Ruud, P. A. (2000), *An Introduction to Classical Econometric Theory*, New York: Oxford University Press. [67]

Saville, D. J., and Wood, G. L. (1991), *Statistical Methods: The Geometric Approach*, New York: Springer-Verlag. [67]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [67]