Principal Component Analysis

Frank Wood

December 8, 2009

This lecture borrows and *quotes* from Joliffe's Principle Component Analysis book. Go buy it!

Principal Component Analysis

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few *retain most of the variation present in* all *of the original variables. [Jolliffe, Pricipal Component Analysis, 2^{nd} edition]*

Data distribution (inputs in regression analysis)

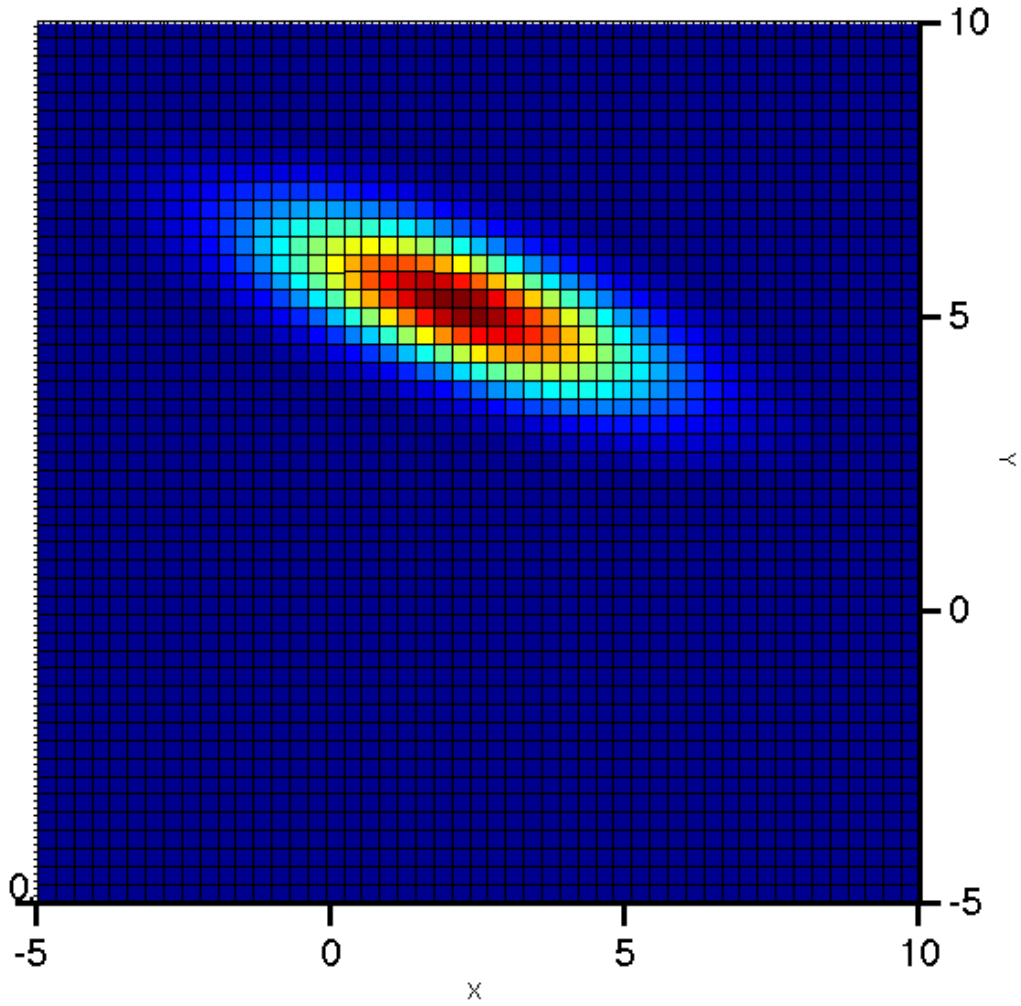

Figure: Gaussian PDF

Uncorrelated projections of principal variation

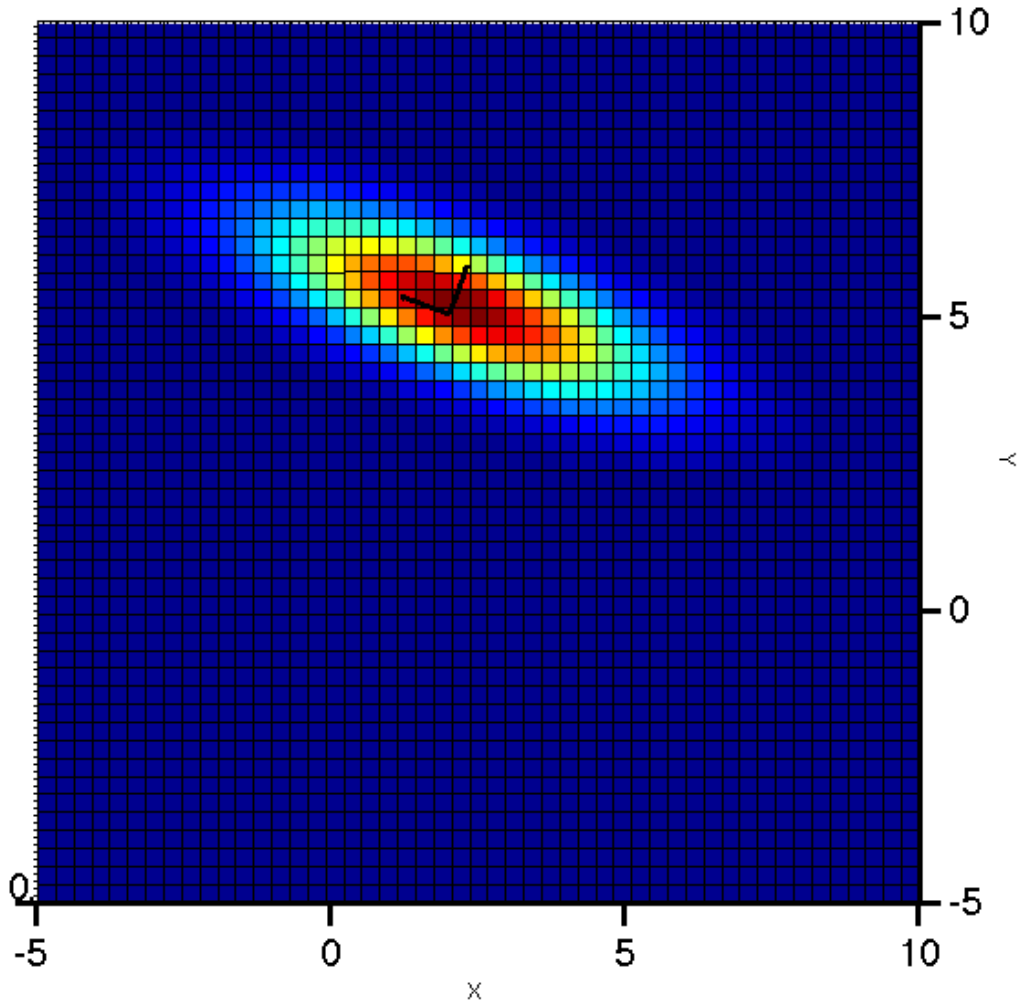

Figure: Gaussian PDF with PC eigenvectors

PCA rotation

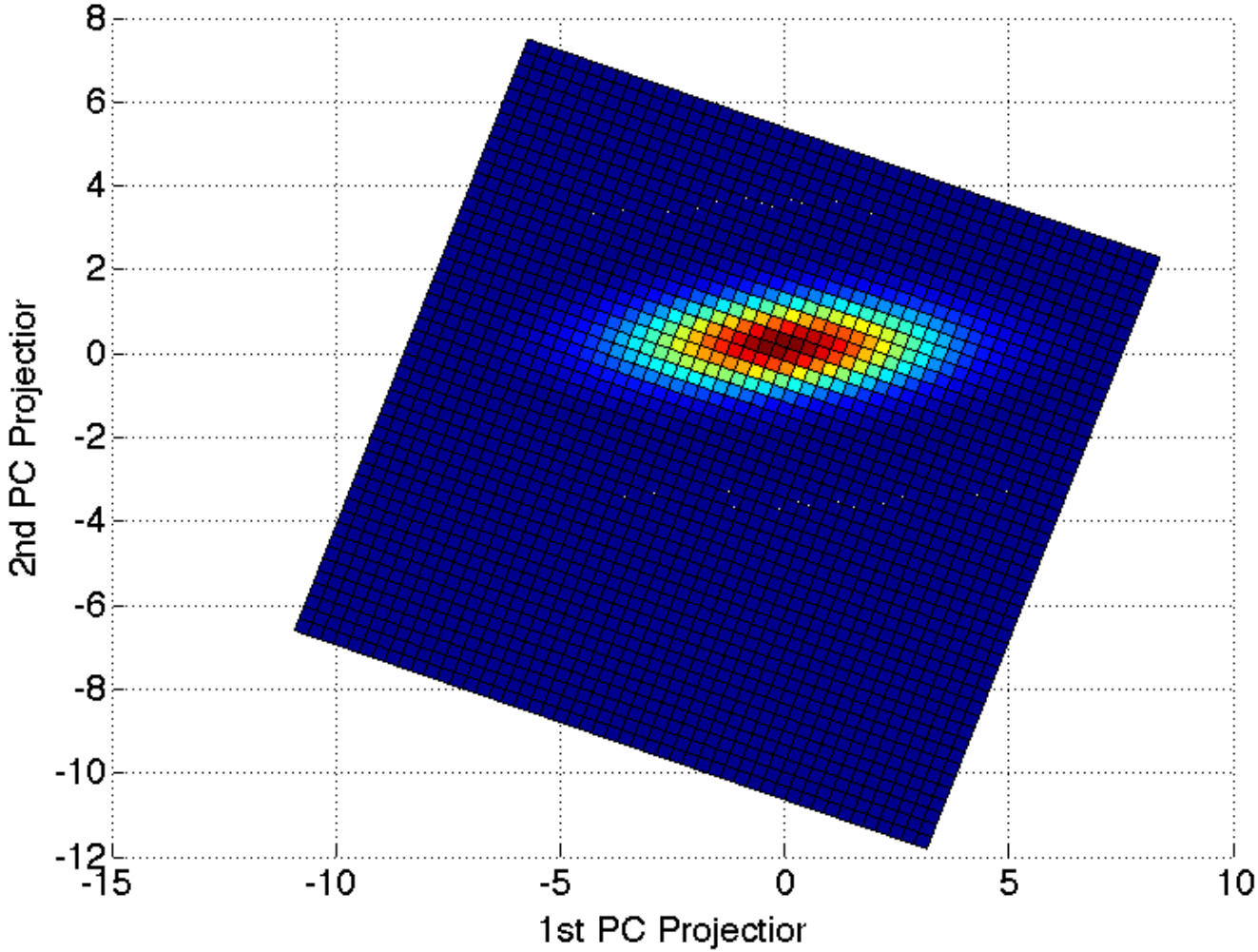

Figure: PCA Projected Gaussian PDF

PCA in a nutshell

Notation

- $\mathbf{x}$ is a vector of p random variables
- $\boldsymbol{\alpha}_k$ is a vector of p constants
- $\boldsymbol{\alpha}_k'\mathbf{x} = \sum_{j=1}^{p} \alpha_{kj} x_j$

Procedural description

- Find linear function of $\mathbf{x}$, $\boldsymbol{\alpha}_1'\mathbf{x}$ with maximum variance.
- Next find another linear function of $\mathbf{x}$, $\boldsymbol{\alpha}_2'\mathbf{x}$, uncorrelated with $\boldsymbol{\alpha}_1'\mathbf{x}$ maximum variance.
- Iterate.

Goal

It is hoped, in general, that most of the variation in $\mathbf{x}$ will be accounted for by m PC's where $m << p$.

Derivation of PCA

Assumption and More Notation

- $\mathbf{\Sigma}$ is the *known* covariance matrix for the random variable $\mathbf{x}$
- Foreshadowing : $\mathbf{\Sigma}$ will be replaced with $\mathbf{S}$, the sample covariance matrix, when $\mathbf{\Sigma}$ is unknown.

Shortcut to solution

- For $k = 1, 2, \ldots, p$ the k^{th} PC is given by $z_k = \boldsymbol{\alpha}_k' \mathbf{x}$ where $\boldsymbol{\alpha}_k$ is an eigenvector of $\mathbf{\Sigma}$ corresponding to its k^{th} largest eigenvalue λ_k.
- If $\boldsymbol{\alpha}_k$ is chosen to have unit length (i.e. $\boldsymbol{\alpha}_k' \boldsymbol{\alpha}_k = 1$) then $\text{Var}(z_k) = \lambda_k$

Derivation of PCA

First Step

- Find $\boldsymbol{\alpha}_k'\mathbf{x}$ that maximizes $\text{Var}(\boldsymbol{\alpha}_k'\mathbf{x}) = \boldsymbol{\alpha}_k'\boldsymbol{\Sigma}\boldsymbol{\alpha}_k$
- Without constraint we could pick a very big $\boldsymbol{\alpha}_k$.
- Choose normalization constraint, namely $\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = 1$ (unit length vector).

Constrained maximization - method of Lagrange multipliers

- To maximize $\boldsymbol{\alpha}_k'\boldsymbol{\Sigma}\boldsymbol{\alpha}_k$ subject to $\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = 1$ we use the technique of Lagrange multipliers. We maximize the function

$$\boldsymbol{\alpha}_k'\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - \lambda(\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k - 1)$$

w.r.t. to $\boldsymbol{\alpha}_k$ by differentiating w.r.t. to $\boldsymbol{\alpha}_k$.

Derivation of PCA

Constrained maximization - method of Lagrange multipliers

- This results in

$$\begin{aligned} \frac{d}{d\boldsymbol{\alpha}_k}\left(\boldsymbol{\alpha}_k'\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - \lambda_k(\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k - 1)\right) &= 0 \\ \boldsymbol{\Sigma}\boldsymbol{\alpha}_k - \lambda_k\boldsymbol{\alpha}_k &= 0 \\ \boldsymbol{\Sigma}\boldsymbol{\alpha}_k &= \lambda_k\boldsymbol{\alpha}_k \end{aligned}$$

- This should be recognizable as an eigenvector equation where $\boldsymbol{\alpha}_k$ is an eigenvector of $\Sigma_b f$ and λ_k is the associated eigenvalue.
- Which eigenvector should we choose?

Derivation of PCA

Constrained maximization - method of Lagrange multipliers

- If we recognize that the quantity to be maximized

$$\boldsymbol{\alpha}_k' \boldsymbol{\Sigma} \boldsymbol{\alpha}_k = \boldsymbol{\alpha}_k' \lambda_k \boldsymbol{\alpha}_k = \lambda_k \boldsymbol{\alpha}_k' \boldsymbol{\alpha}_k = \lambda_k$$

  then we should choose λ_k to be as big as possible. So, calling λ_1 the largest eigenvector of Σ and $\boldsymbol{\alpha}_1$ the corresponding eigenvector then the solution to

$$\boldsymbol{\Sigma} \boldsymbol{\alpha}_1 \quad = \quad \lambda_1 \boldsymbol{\alpha}_1$$

  is the 1^{st} principal component of $\mathbf{x}$.
- In general $\boldsymbol{\alpha}_k$ will be the k^{th} PC of $\mathbf{x}$ and $\text{Var}(\boldsymbol{\alpha}'\mathbf{x}) = \lambda_k$
- We will demonstrate this for $k = 2$, $k > 2$ is more involved but similar.

Derivation of PCA

Constrained maximization - more constraints

- The second PC, $\boldsymbol{\alpha}_2\mathbf{x}$ maximizes $\boldsymbol{\alpha}_2\boldsymbol{\Sigma}\boldsymbol{\alpha}_2$ subject to being uncorrelated with $\boldsymbol{\alpha}_1\mathbf{x}$.
- The uncorrelation constraint can be expressed using any of these equations

$$\begin{aligned}\text{cov}(\boldsymbol{\alpha}_1'\mathbf{x}, \boldsymbol{\alpha}_2'\mathbf{x}) &= \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2'\lambda_1\boldsymbol{\alpha}_1' \\ &= \lambda_1\boldsymbol{\alpha}_2'\boldsymbol{\alpha} = \lambda_1\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 = 0\end{aligned}$$

- Of these, if we choose the last we can write an Langrangian to maximize $\boldsymbol{\alpha}_2$

$$\boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2(\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 - 1) - \phi\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1$$

Derivation of PCA

Constrained maximization - more constraints

- Differentiation of this quantity w.r.t. $\boldsymbol{\alpha}_2$ (and setting the result equal to zero) yields

$$\frac{d}{d\boldsymbol{\alpha}_2}\left(\boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2(\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 - 1) - \phi\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1\right) = 0$$
$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1 = 0$$

- If we left multiply $\boldsymbol{\alpha}_1$ into this expression

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 0$$
$$0 - 0 - \phi 1 = 0$$

then we can see that ϕ must be zero and that when this is true that we are left with

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha}_2 = 0$$

Derivation of PCA

Clearly

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha}_2 = 0$$

is another eigenvalue equation and the same strategy of choosing α_2 to be the eigenvector associated with the second largest eigenvalue yields the second PC of $\mathbf{x}$, namely $\boldsymbol{\alpha}_2'\mathbf{x}$.

This process can be repeated for $k = 1 \ldots p$ yielding up to p different eigenvectors of $\boldsymbol{\Sigma}$ along with the corresponding eigenvalues $\lambda_1, \ldots \lambda_p$.

Furthermore, the variance of each of the PC's are given by

$$\text{Var}[\boldsymbol{\alpha}_k'\mathbf{x}] = \lambda_k, \qquad k = 1, 2, \ldots, p$$

Properties of PCA

For any integer $q, 1 \leq q \leq p$, consider the orthonormal linear transformation

$$\mathbf{y} = \mathbf{B}'\mathbf{x}$$

where $\mathbf{y}$ is a q-element vector and $\mathbf{B}'$ is a $q \times p$ matrix, and let $\mathbf{\Sigma}_y = \mathbf{B}'\mathbf{\Sigma}\mathbf{B}$ be the variance-covariance matrix for $\mathbf{y}$. Then the trace of $\mathbf{\Sigma}_y$, denoted $\text{tr}(\mathbf{\Sigma}_y)$, is maximized by taking $\mathbf{B} = \mathbf{A}_q$, where $\mathbf{A}_q$ consists of the first q columns of $\mathbf{A}$.

What this means is that if you want to choose a lower dimensional projection of $\mathbf{x}$, the choice of $\mathbf{B}$ described here is probably a good one. It maximizes the (retained) variance of the resulting variables.

In fact, since the projections are uncorrelated, the percentage of variance accounted for by retaining the first q PC's is given by

$$\frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k} \times 100$$

PCA using the sample covariance matrix

If we recall that the sample covariance matrix (an unbiased estimator for the covariance matrix of $\mathbf{x}$) is given by

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}'\mathbf{X}$$

where $\mathbf{X}$ is a $(n \times p)$ matrix with $(i,j)th$ element $(x_{ij} - \bar{x}_j)$ (in other words, $\mathbf{X}$ is a zero mean design matrix).

We construct the matrix $\mathbf{A}$ by combining the p eigenvectors of $\mathbf{S}$ (or eigenvectors of $\mathbf{X}'\mathbf{X}$ – they're the same) then we can define a matrix of PC scores

$$\mathbf{Z} = \mathbf{X}\mathbf{A}$$

Of course, if we instead form $\mathbf{Z}$ by selecting the q eigenvectors corresponding to the q largest eigenvalues of $\mathbf{S}$ when forming $\mathbf{A}$ then we can achieve an "optimal" (in some senses) q-dimensional projection of $\mathbf{x}$.

Computing the PCA loading matrix

Given the sample covariance matrix

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}'\mathbf{X}$$

the most straightforward way of computing the PCA loading matrix is to utilize the singular value decomposition of $\mathbf{S} = \mathbf{A}'\mathbf{\Lambda}\mathbf{A}$ where $\mathbf{A}$ is a matrix consisting of the eigenvectors of $\mathbf{S}$ and $\mathbf{\Lambda}$ is a diagonal matrix whose diagonal elements are the eigenvalues corresponding to each eigenvector.

Creating a reduced dimensionality projection of $\mathbf{X}$ is accomplished by selecting the q largest eigenvalues in $\mathbf{\Lambda}$ and retaining the q corresponding eigenvectors from $\mathbf{A}$

Sample Covariance Matrix PCA

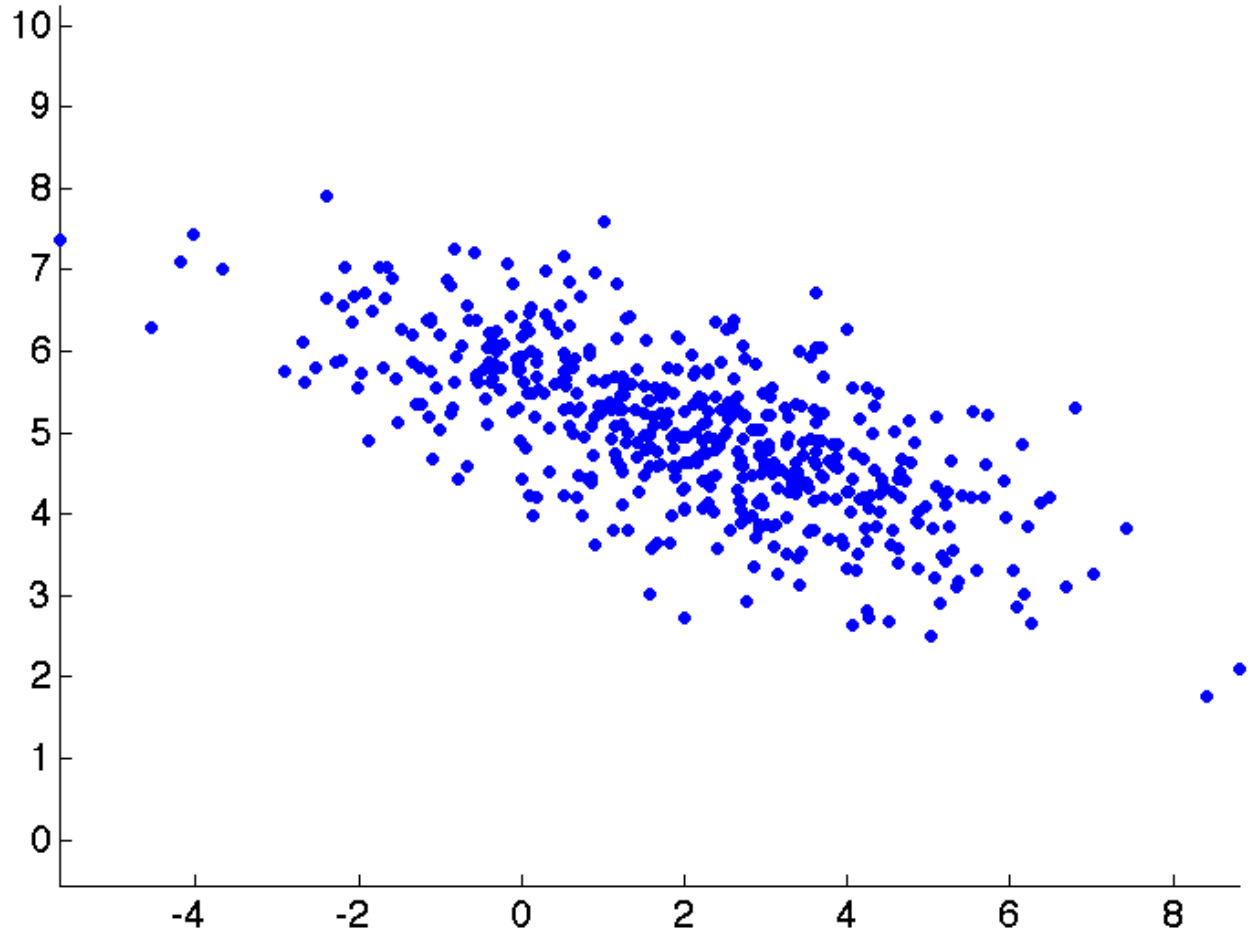

Figure: Gaussian Samples

Sample Covariance Matrix PCA

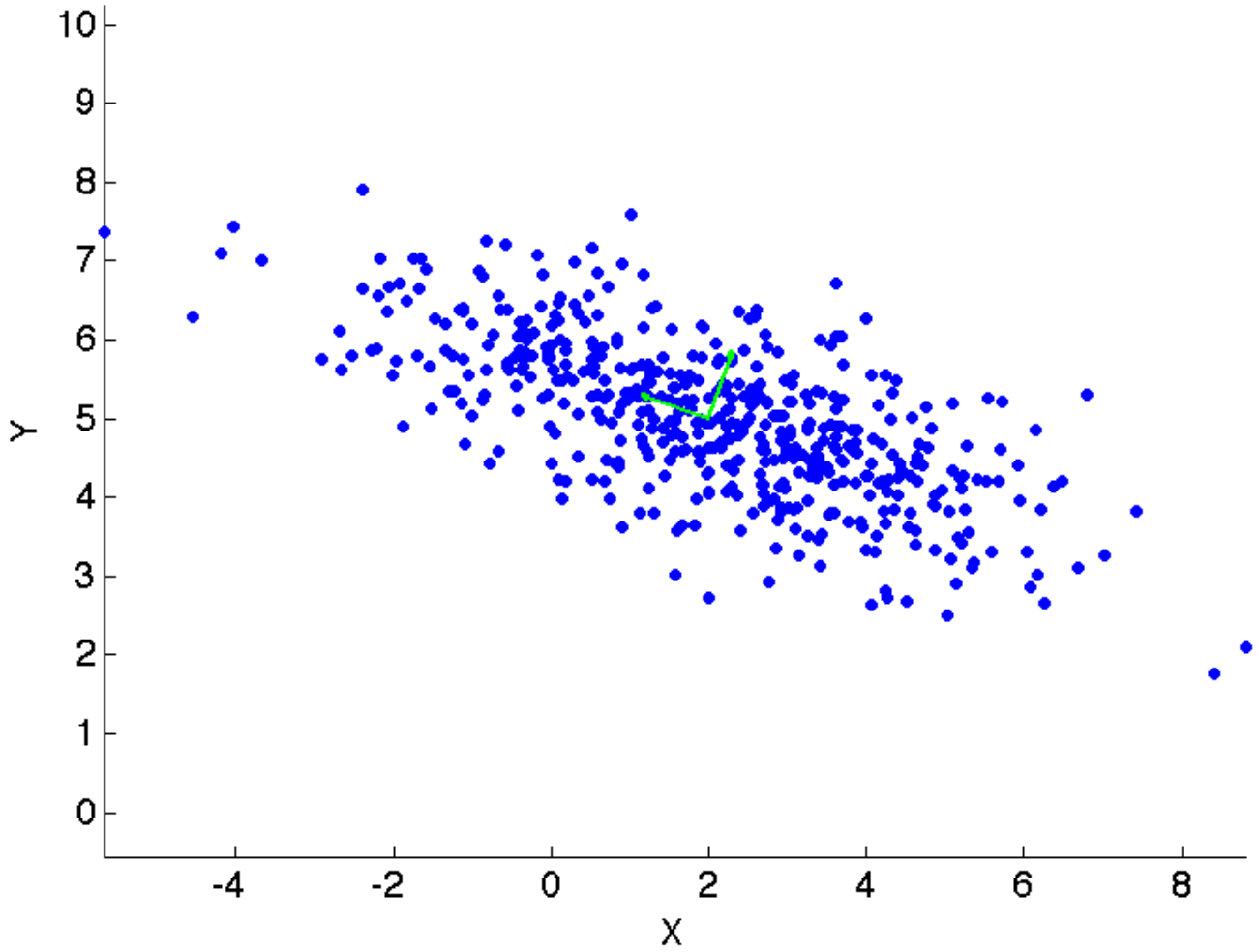

Figure: Gaussian Samples with eigenvectors of sample covariance matrix

Sample Covariance Matrix PCA

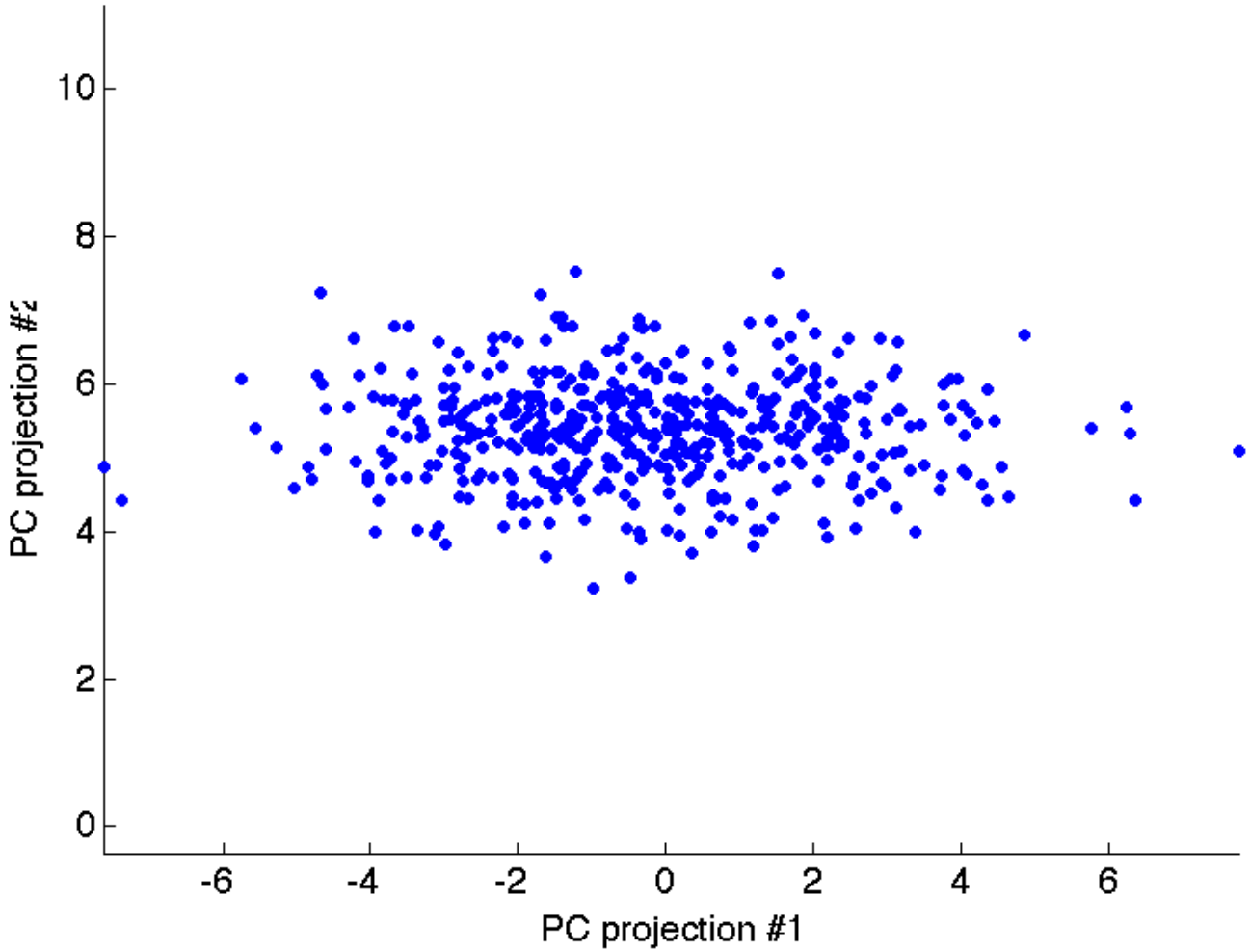

Figure: PC projected samples

Sample Covariance Matrix PCA

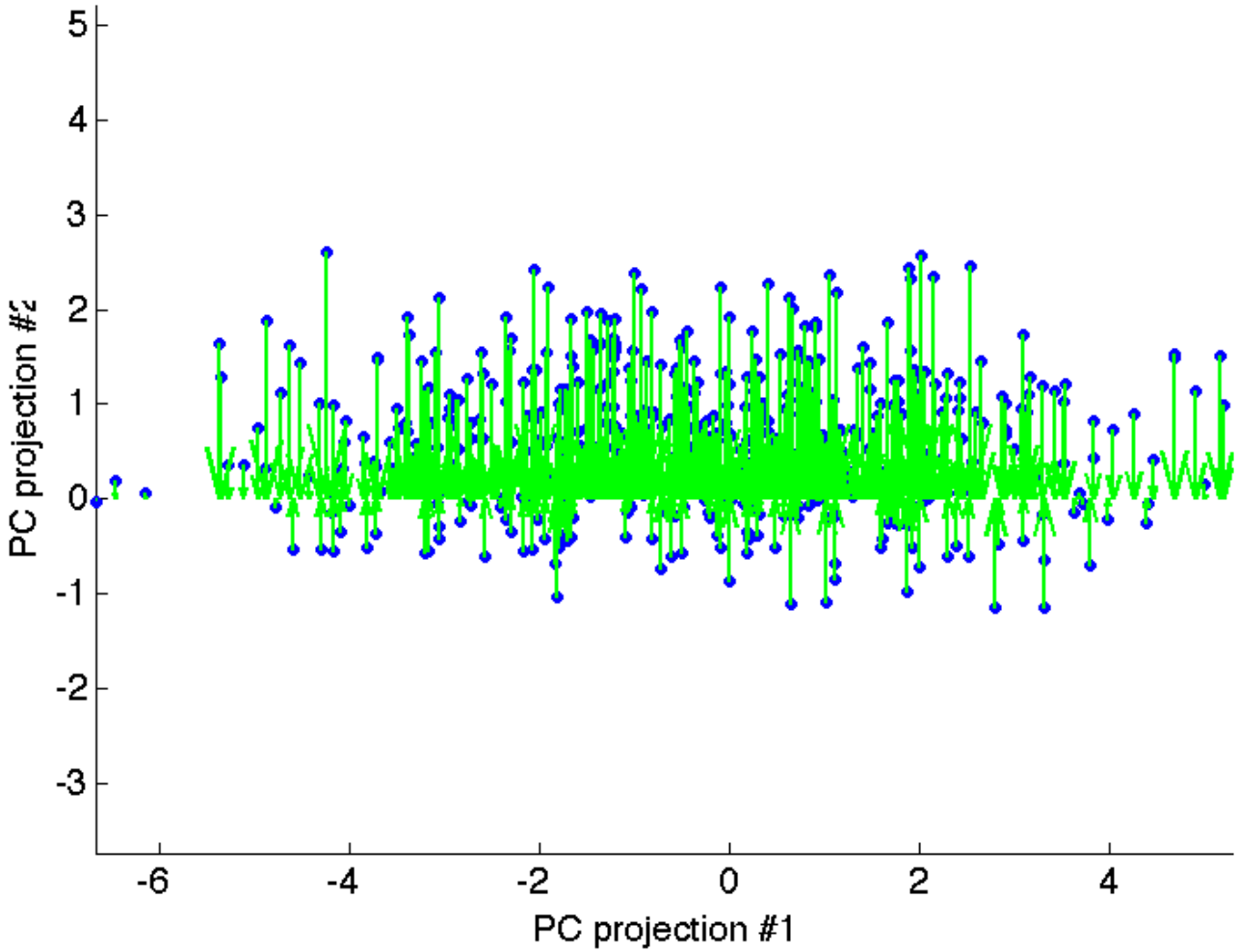

Figure: PC dimensionality reduction step

Sample Covariance Matrix PCA

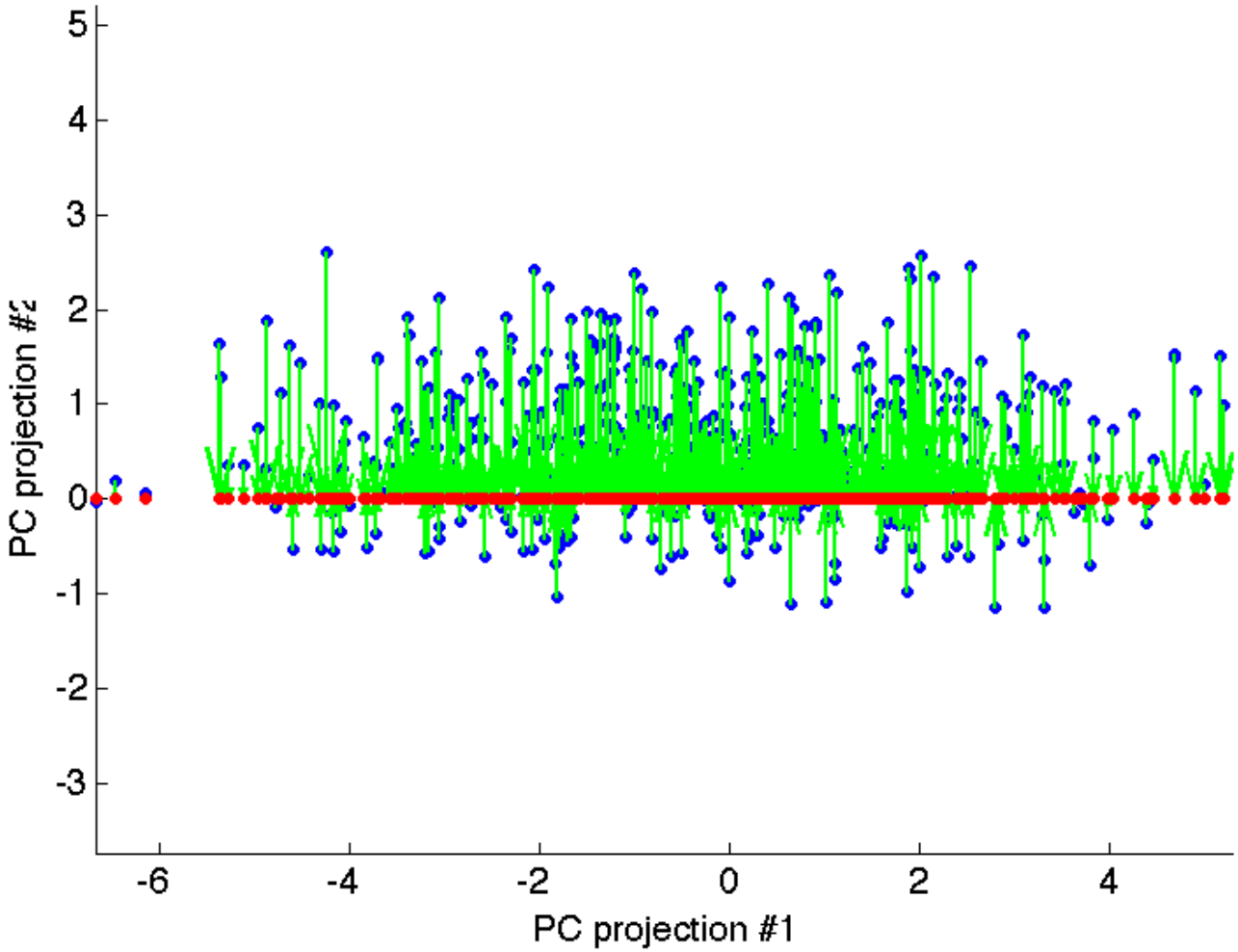

Figure: PC dimensionality reduction step

PCA in linear regression

PCA is useful in linear regression in several ways

- Identification and elimination of multicolinearities in the data.
- Reduction in the dimension of the input space leading to fewer parameters and "easier" regression.
- Related to the last point, the variance of the regression coefficient estimator is minimized by the PCA choice of basis.

We will consider the following example.

- $\mathbf{x} \sim \mathrm{N}\left([2\ 5], \begin{bmatrix} 4.5 & -1.5 \\ -1.5 & 1.0 \end{bmatrix}\right)$
- $\mathbf{y} = \mathbf{X}[-1\ 2]'$ when no colinearities are present (no noise)
- $x_{i3} = .8x_{i1} + .5x_{i2}$ *imposed* colinearity

Noiseless Linear Relationship with No Colinearity

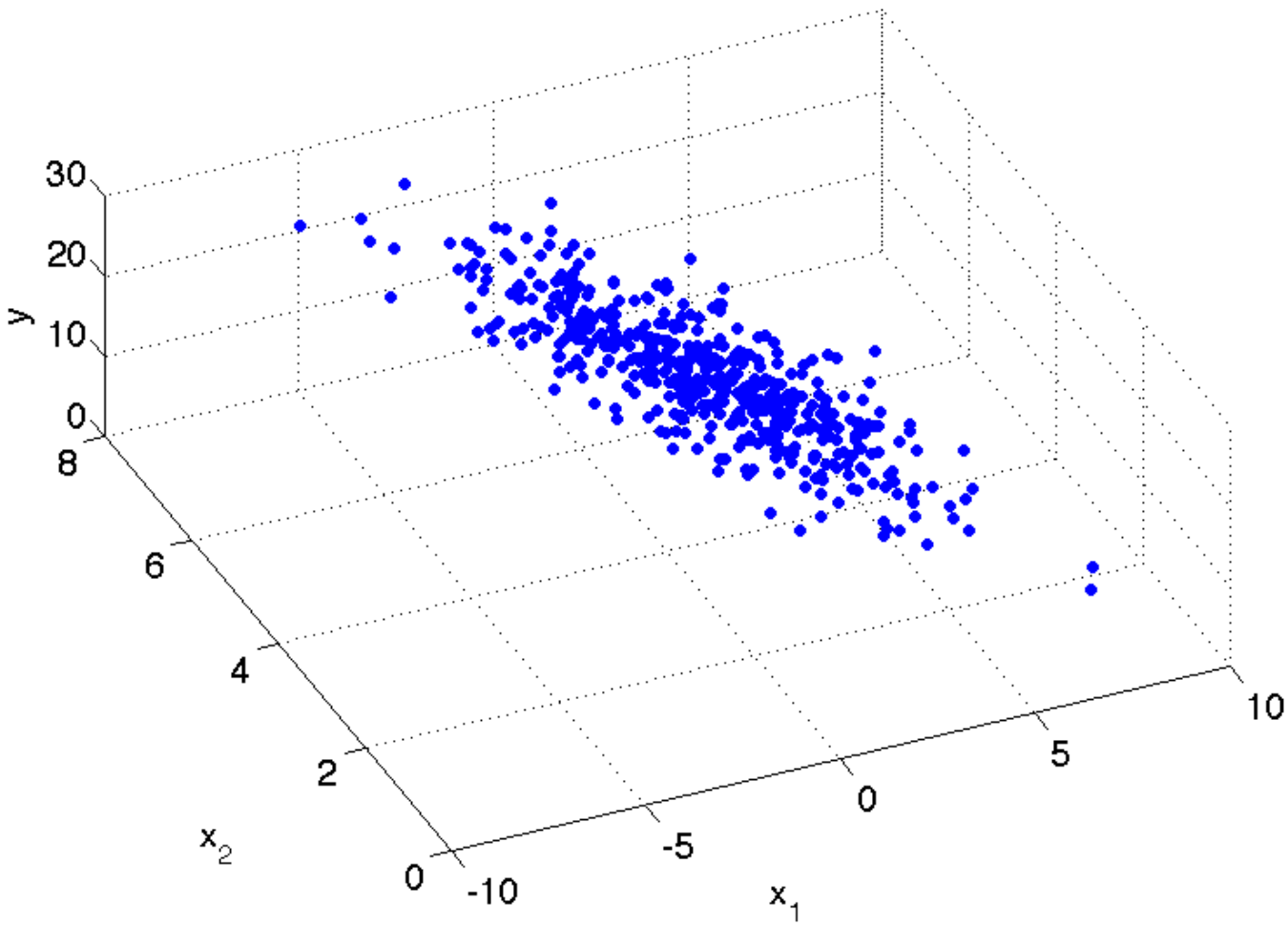

Figure: $\mathbf{y} = \mathbf{x}[-1\ 2]' + 5, \mathbf{x} \sim \mathrm{N}([2\ 5], \begin{bmatrix} 4.5 & -1.5 \\ -1.5 & 1.0 \end{bmatrix})$

Noiseless Planar Relationship

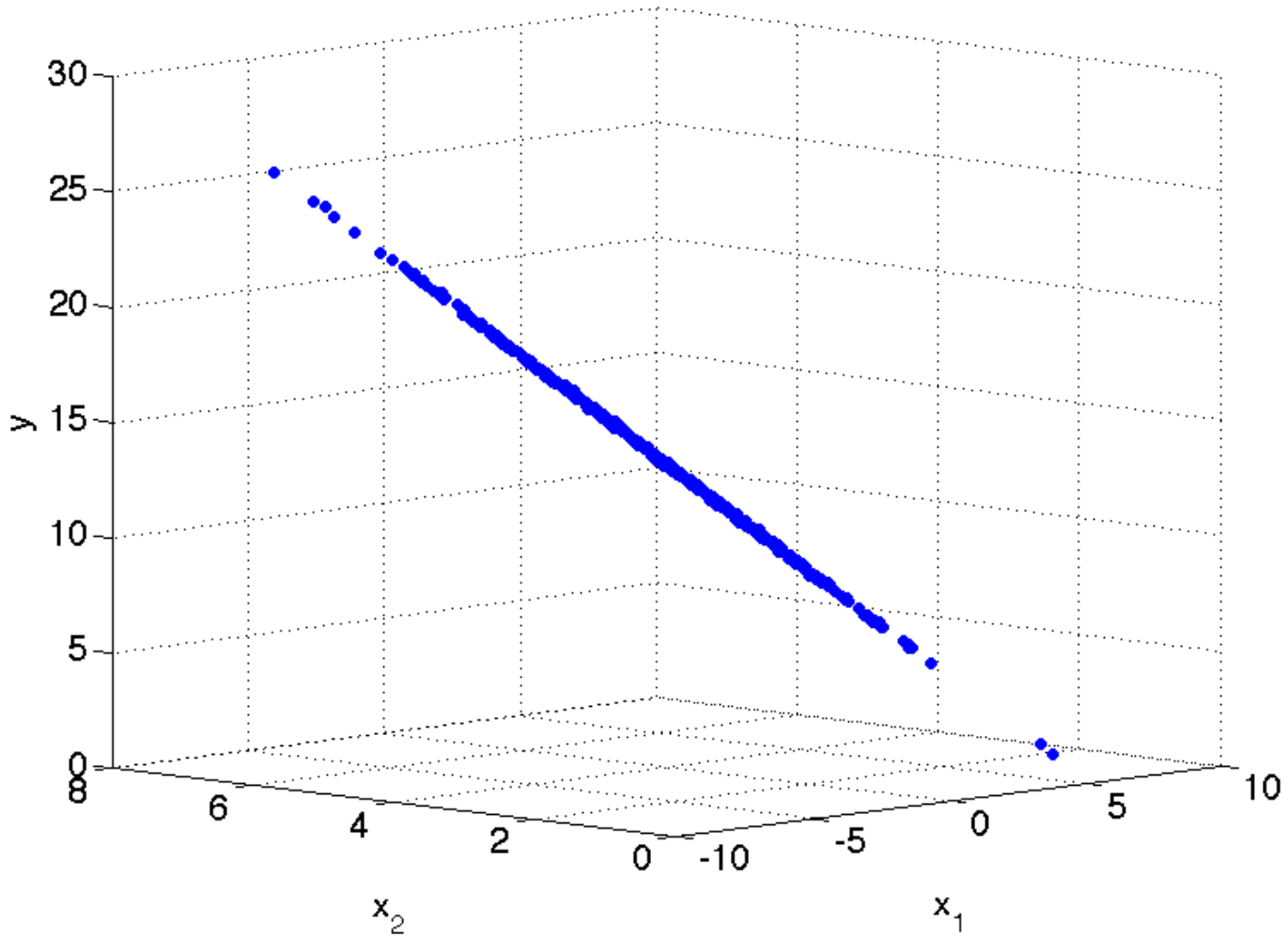

Figure: $\mathbf{y} = \mathbf{x}[-1\ 2]' + 5, \mathbf{x} \sim \mathrm{N}([2\ 5], \begin{bmatrix} 4.5 & -1.5 \\ -1.5 & 1.0 \end{bmatrix})$

Projection of colinear data

The figures before showed the data without the third colinear design matrix column. Plotting such data is not possible, but it's colinearity is obvious by design.

When PCA is applied to the design matrix of rank q less than p the number of positive eigenvalues discovered is equal to q the true rank of the design matrix.

If the number of PC's retained is larger than q (and the data is perfectly colinear, etc.) *all* of the variance of the data is retained in the low dimensional projection.

In this example, when PCA is run on the design matrix of rank 2, the resulting projection back into two dimensions has exactly the same distribution as before.

Projection of colinear data

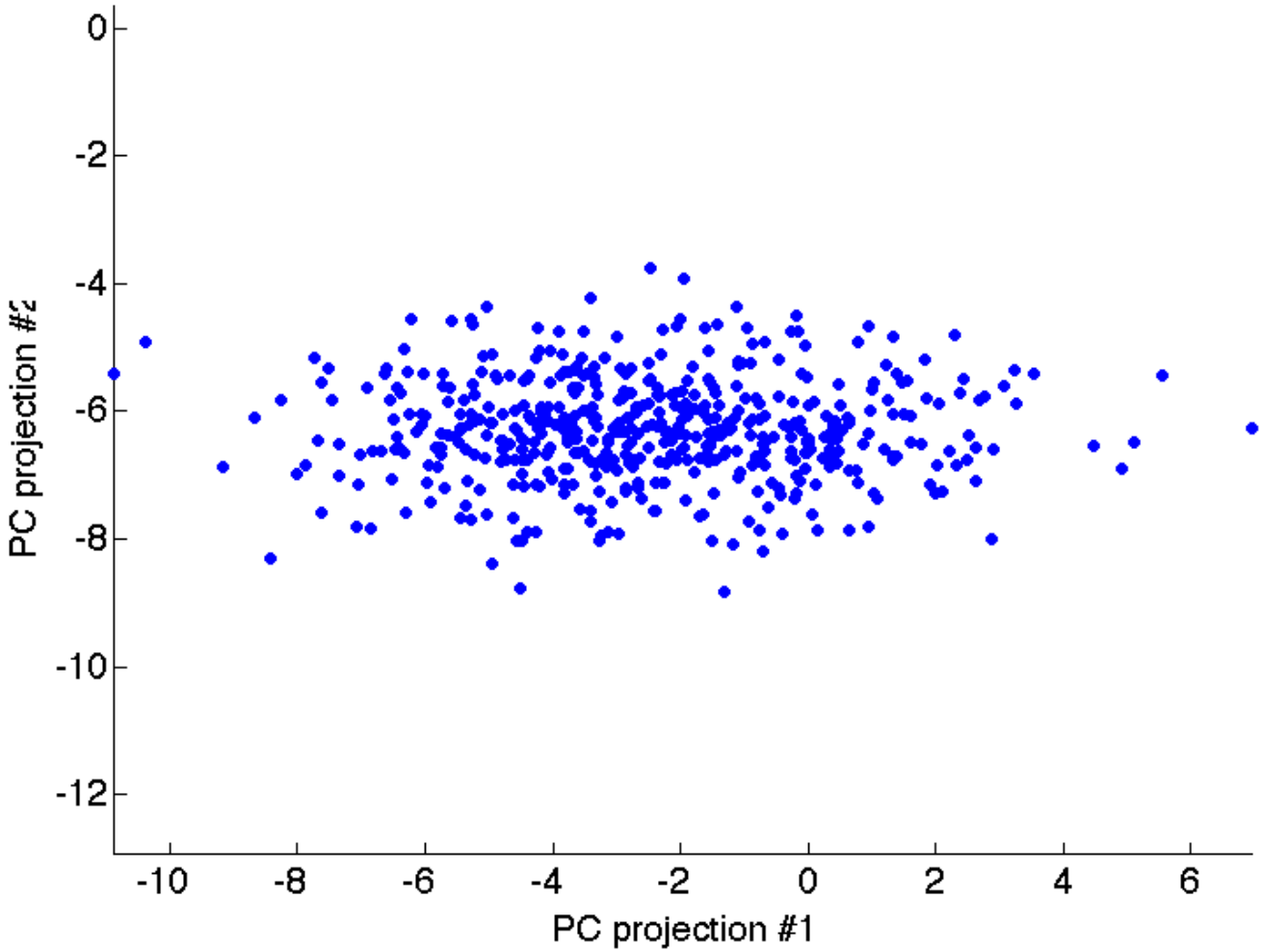

Figure: Projection of multi-colinear data onto first two PC's

Reduction in regression coefficient estimator variance

If we take the standard regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

And consider instead the PCA rotation of $\mathbf{X}$ given by

$$\mathbf{Z} = \mathbf{Z}\mathbf{A}$$

then we can rewrite the regression model in terms of the PC's

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}.$$

We can also consider the reduced model

$$\mathbf{y} = \mathbf{Z}_q\boldsymbol{\gamma}_q + \boldsymbol{\epsilon}_q$$

where only the first q PC's are retained.

Reduction in regression coefficient estimator variance

If we rewrite the regression relation as

$$\mathbf{y} = \mathbf{Z}\gamma + \epsilon.$$

Then we can, because $\mathbf{A}$ is orthogonal, rewrite

$$\mathbf{X}\beta = \mathbf{X}\mathbf{A}\mathbf{A}'\beta = \mathbf{Z}\gamma$$

where $\gamma = \mathbf{A}'\beta$.

Clearly using least squares (or ML) to learn $\hat{\beta} = \mathbf{A}\hat{\gamma}$ is equivalent to learning $\hat{\beta}$ directly.

And, like usual,

$$\hat{\gamma} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

so $\hat{\beta} = \mathbf{A}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$

Reduction in regression coefficient estimator variance

Without derivation we note that the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \sum_{k=1}^{p} l_k^{-1} \mathbf{a}_k \mathbf{a}_k'$$

where l_k is the k^{th} largest eigenvalue of $\mathbf{X}'\mathbf{X}$, $\mathbf{a}_k$ is the k^{th} column of $\mathbf{A}$, and σ^2 is the observation noise variance, i.e. $\boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}, \sigma^2\mathbf{I})$

This sheds light on how multicolinearities produce large variances for the elements of $\hat{\boldsymbol{\beta}}$. If an eigenvector l_k is small then the resulting variance of the estimator will be large.

Reduction in regression coefficient estimator variance

One way to avoid this is to ignore those PC's that are associated with small eigenvalues, namely, use biased estimator

$$\tilde{\boldsymbol{\beta}} = \sum_{k=1}^{m} l_k^{-1} \mathbf{a}_k \mathbf{a}_k' \mathbf{X}' \mathbf{y}$$

where $l_{1:m}$ are the large eigenvalues of $\mathbf{X}'\mathbf{X}$ and $l_{m+1:p}$ are the small.

$$\text{Var}(\tilde{\boldsymbol{\beta}}) = \sigma^2 \sum_{k=1}^{m} l_k^{-1} \mathbf{a}_k \mathbf{a}_k'$$

This is a biased estimator, but, since the variance of this estimator is smaller it is possible that this could be an advantage.

Homework: find the bias of this estimator. Hint: use the spectral decomposition of $\mathbf{X}'\mathbf{X}$.

Problems with PCA

PCA is not without its problems and limitations

- PCA assumes approximate normality of the input space distribution
  - PCA may still be able to produce a "good" low dimensional projection of the data even if the data isn't normally distributed
- PCA may "fail" if the data lies on a "complicated" manifold
- PCA assumes that the input data is real and continuous.
- Extensions to consider
  - Collins et al, A generalization of principal components analysis to the exponential family.
  - Hyv
    "arinen, A. and Oja, E., Independent component analysis: algorithms and applications
  - ISOMAP, LLE, Maximum variance unfolding, etc.

Non-normal data

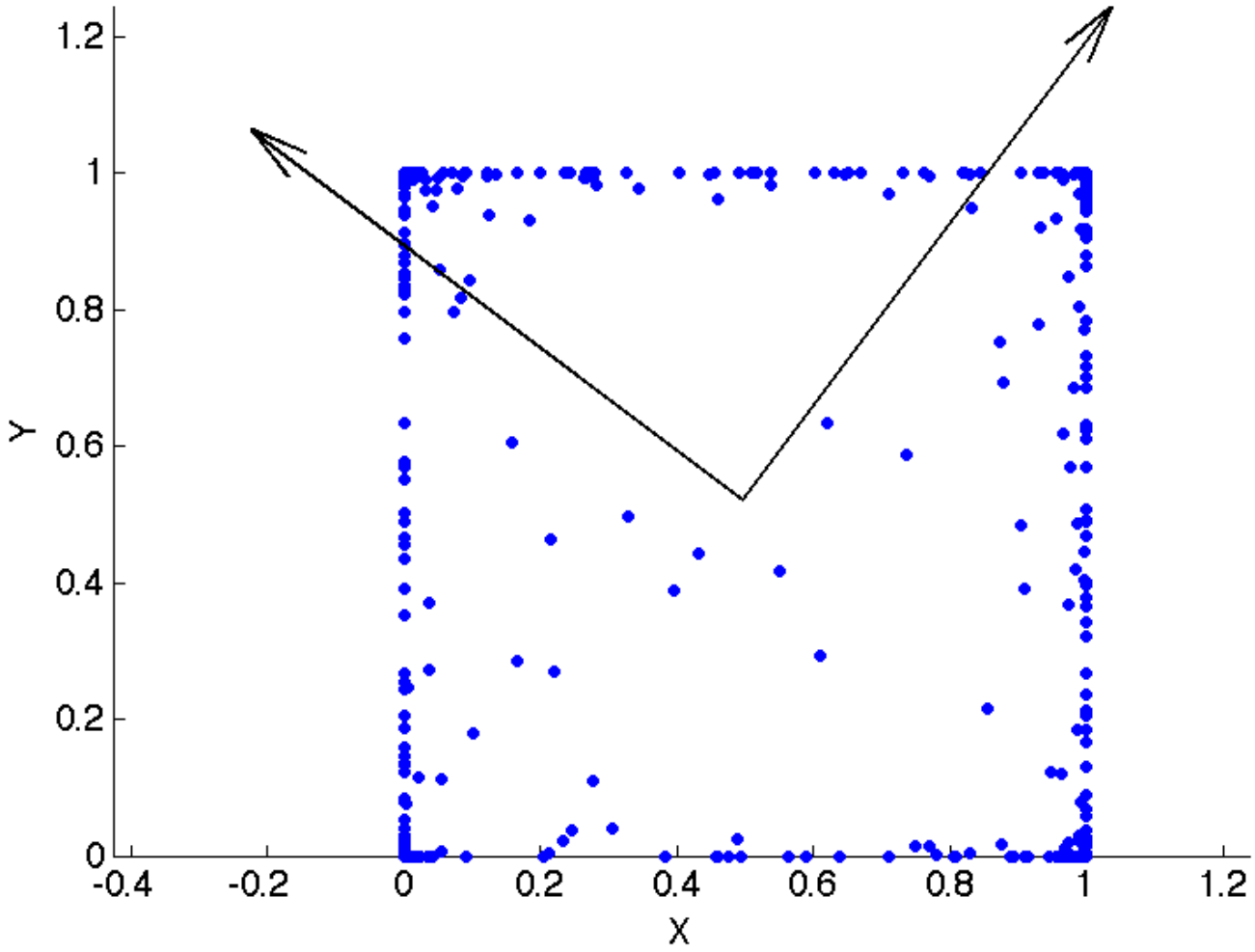

Figure: 2d Beta(.1, .1) Samples with PC's

Non-normal data

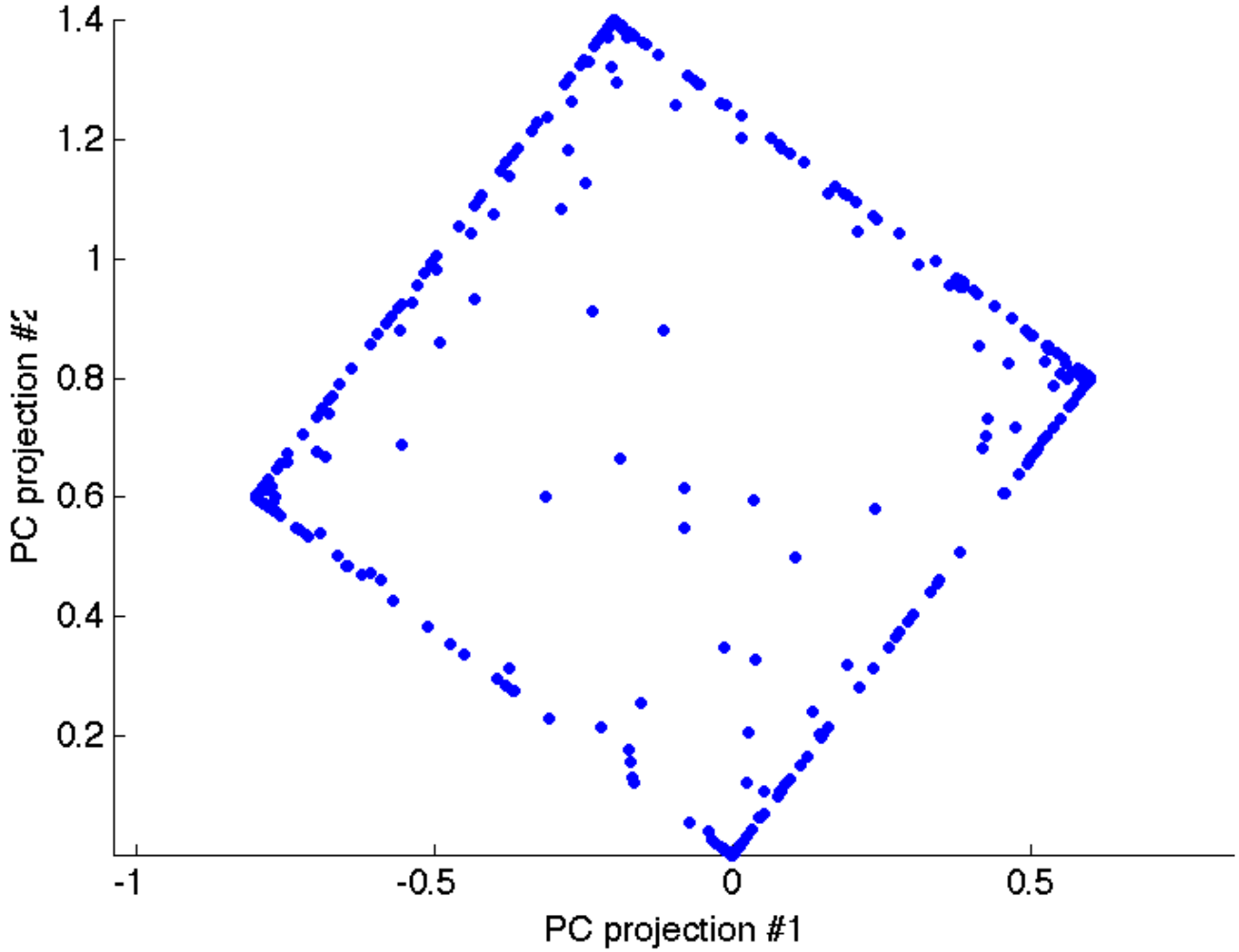

Figure: PCA Projected