Principal Component Analysis.

Linear algebra is one of the great superpowers! A lot of what we do in data science, indeed, a lot of mathematical modelling is either a direct example of linear algebra, or, something more complicated we manage to solve because we can break it up into bits that look like linear algebra. Linear algebra isn't difficult, at the start it just looks like a clever way to organise some complicated calculations, you might say that mathematics is always just a clever way to organise some calculations. There is, however, a deep idea that emerges in linear algebra which we will touch on here and exploit: matrices can be thought of in terms of their components or in terms of their eigenvectors and eigenvalues, what is sometimes called their *spectral content*; we will use this to find diagonal structures in the matrices which we can use to find structure in data.

In this section we are going to look at principal component analysis (PCA). You will actually encounter PCA twice, in this unit and when you are learning about visualization. This reflects the importance of PCA, from a visualization point-of-view because it is often the first step in any examination of data and from a AI point-of-view because it is an early demonstration of the idea that data usually has a structure and this is why AI works as well as it does! PCA is an unsupervised technique but a good supervised learning algorithm will always have a hidden unsupervised aspect, it discovers the structure of data on the way to learning how you want it classified. As well as being useful example of unsupervised learning, it reminds us a bit about linear algebra and it teaches us a bit about how to think about data. However, although we won't go into this explicitly, keep in mind that many interesting aspects of neural networks are understandable as non-linear generalizations of PCA! Anyway, I hope you won't mind this small piece of repetition between units, it is worth doing twice and I am sure it will be done with different emphasis in both cases, I am sure the other lecturer will also show more skill in teaching the material!

The first thing is *unsupervised learning*; by unsupervised learning we mean, broadly, we don't start with a set of data points and some labels and try to learn how to map from data to labels, that's supervised learning; in unsupervised learning we have some data points and we try to decide what structure the points have, what is interesting about them, before actually thinking about labelling. Old fashioned text books make a big deal about the distinction between supervised and unsupervised learning; often, in real

life, we find that distinction is not so clear; a neural network, for example, is typically trained on labelled data, so they are exampled of supervised systems, but often earlier layers do something more akin to unsupervised learning, breaking input down by its intrinsic structure, all the better to be classified by later layer.

I am overly inclined to talk about things in terms of what the brain does, but the brain clearly performs some mixture of unsupervised and supervised learning. A baby, when it first is learning to see, when its visual system is wiring itself up, is being presented with visual data, this data has structure, there are edges, the edges surround objects, objects appear in front of each other in a sort of consistient way, they have shading; the same object can be large or small according to how far away or how close it is. This the baby learns before it starts to learn to recognize different things; that is the sort of learning we call unsupervised.

Some linear algebra!

I don't want to spend ages reminding you how linear algebra works; you can look this up. However, it might be useful to have a quick recap. There are lots of ways to think about what matrices do, in fact, it is a clue to what a powerful concept they are that they can be introduced in some many different ways!

Here is one version! Consider the simulateous equations:

$$\begin{aligned} x + y &= 3 \\ x - y &= 1 \end{aligned} \tag{1}$$

We can solve this easily by adding the two equation, this give $2x = 4$ or $x = 2$, substituting that back in gives $y = 1$ and it is easy to check that this solve both equations. The geometrical explanation is that both equations are equations for lines and the point $(2, 1)$ is the unique point that lies on both lines solving both equations.

We can rewrite this in matrix form:

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \tag{2}$$

which is in the matrix form

$$A\mathbf{x} = \mathbf{y} \tag{3}$$

Provided we can find an inverse matrix A^{-1} so that $A^{-1}A = \mathbf{1}$ then we can solve the equation:

$$A^{-1}A\mathbf{x} = A^{-1}\mathbf{y} \tag{4}$$

or

$$\mathbf{x} = A^{-1}\mathbf{y} \tag{5}$$

In fact if

$$A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{6}$$

as above, then

$$A^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{7}$$

or

$$\begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \tag{8}$$

So what did that palaver with the matrix get us? In a way it deskilled the solving of the simultaneous equation; the first time around, when we solved it without using matrices we began by adding the two equations so the y terms cancelled and we got a value for x. This didn't require much skill since it was a very easy example, but you can see that some thinking is required. The matrix method, in contrast, only relied on the step I sort of skipped, inverting the matrix; if you write the equation in terms of matrices then you can solve them by inverting the matrix. In fact, any method of solving the matrix is equivalent to inverting the matrix, so solving the equation is exactly as hard as inverting the matrix. We have, over the last 100 years put a lot of effort into finding ways to invert matrices, it is a long-winded calculation, but one we are very good at, and by "we" I mean linear algebra packages.

Not all matrices can be inverted: this matrix

$$A = \begin{pmatrix} 2 & 3 \\ -4 & -6 \end{pmatrix} \tag{9}$$

for example, has no inverse. That's ok though because not all simultaneous equations can be solved, for example

$$\begin{aligned} 2x + 3y &= 3 \\ -4x - 6y &= 1 \end{aligned} \tag{10}$$

has no solution because the two lines, the $2x+3y = 3$ line and the $-4x-6y = 1$ line are parallel. In fact, I can tell straight away that the matrix can't be

inverted by working out its determinant. The determinant is a quantity associated with matrices, in the 2×2 case it has a simple formula:

$$\det \begin{pmatrix} a & c \\ d & b \end{pmatrix} = ab - cd \tag{11}$$

So

$$\det \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = -2 \tag{12}$$

and

$$\det \begin{pmatrix} 2 & 3 \\ -4 & -6 \end{pmatrix} = 0 \tag{13}$$

The rule is that a matrix has an inverse if its determinant isn't zero. The formula for working out the determinant of bigger matrices is tedious to explain, but easily programmed, we can easily work out determinants, though for anything bigger than 2×2 or certainly 3×3 it's best to get a computer to do it for you!

Eigenvectors and eigenvalues

Ok so this bit is the special secret bit. Matrices can be thought of as being made up of their eigenvectors. The word *eigen* means equal in German and an eigenvector for a matrix is a direction that isn't changed by the matrix, so it satisfies:

$$A\mathbf{e} = \lambda \mathbf{e} \tag{14}$$

In other words, multiplying the eigenvector by the matrix can make it longer or shorter, according to whether λ is bigger or smaller than one, but it stays pointing in the same direction. It isn't clear yet why this is a big deal, but it will turn out to be a nice way of thinking how matrices work and we'll use it to do PCA!

How do you find the eigenvectors? Well this is a lovely piece of using mathematics in a kind of wrong way around trick. We want

$$A\mathbf{e} = \lambda \mathbf{e} \tag{15}$$

so we can write that as

$$(A - \lambda \mathbf{1})\, \mathbf{e} = \mathbf{0} \tag{16}$$

In our simple example from earlier imagine we want

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \lambda \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \tag{17}$$

then we rewrite this as

$$\begin{pmatrix} 1-\lambda & 1 \\ 1 & -1-\lambda \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{18}$$

Back though to the general version

$$(A - \lambda \mathbf{1})\,\mathbf{e} = \mathbf{0} \tag{19}$$

for convenience write $A_\lambda = A - \lambda \mathbf{1}$ so our equation becomes

$$A_\lambda \mathbf{e} = \mathbf{0} \tag{20}$$

Now if we invert the matrix A_λ everything goes wrong:

$$\mathbf{e} = A_\lambda^{-1}\mathbf{0} = \mathbf{0} \tag{21}$$

so for the eigenvector to be something interesting, that is not just zero, the matrix A_λ must have no inverse! Luckily we know how to check that, it means

$$\det A_\lambda = 0 \tag{22}$$

which we'll see in a second is just an equation for λ; this equation is called the *characteristic equation*.

Let's try this for a simple example, I amn't going to use the matrix above, just because it turns out to have $\sqrt{2}$'s in its eigenvalues, it isn't hard but lets keep things as straightforward as possible and instead lets look at the example:

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \tag{23}$$

so

$$A_\lambda = \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix} \tag{24}$$

then

$$\det A_\lambda = (2-\lambda)^2 - 1 = \lambda^2 - 4\lambda + 3 \tag{25}$$

so the characteristic equation is

$$\lambda^2 - 4\lambda + 3 = 0 \tag{26}$$

which has solutions $\lambda = 3$ and $\lambda = 1$.

It turns out once you know the eigenvalues you can easily work out the eigenvectors by substituting back into the original equation. You can look up how to do this, it isn't hard. In this case for $\lambda = 3$ the eigenvector is

$$\mathbf{e} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{27}$$

and for $\lambda = 1$

$$\mathbf{e} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{28}$$

There is a small subtlety here which is basically that an eigenvector is a direction not a vector, if $\mathbf{e}$ is an eigenvector then so is $\mu\mathbf{e}$ for any non-zero μ, this hinges on the linearity of matrices, say $A\mathbf{e} = \lambda\mathbf{e}$ then $A(\mu\mathbf{e}) = \mu A\mathbf{e} = \mu\lambda\mathbf{e} = \lambda(\mu\mathbf{e})$, so if $\mathbf{e}$ is an eigenvector, so is $\mu\mathbf{e}$.

Diagonalization

What good are eigenvectors? Well, as I said before, you can think of matrices as being made up of their eigenvectors; I don't want to go into this in detail, there are a lot of details, but the example I want to look at is diagonalization, this is a related sort of idea and then one we need for doing PCA!

To talk about diagonalization, I am going to assume the matrix is symmetric. This isn't required, there are diagonalizable non-symmetric matrices, but the example we need for PCA is symmetric and the symmetric case lacks all the caveats that we need for the more general case, by focusing on diagonalizable we get to avoid some "as long as blah blah blah" type statements.

To remind you about symmetry, the transpose of a matrix involves flipping around the diagonal, so if $A = [a_{ij}]$ then $A^T = [a_{ji}]$, for 2×2 is

$$A = \begin{pmatrix} a & c \\ d & b \end{pmatrix} \tag{29}$$

then

$$A^T = \begin{pmatrix} a & d \\ c & b \end{pmatrix} \tag{30}$$

A matrix is symmetric if $A = A^T$ so

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix} \tag{31}$$

is symmetric,

$$A = \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix} \tag{32}$$

is not.

Now, what do we mean by diagonalization, it basically means that we can rewrite a symmetric matrix in the form

$$A = PDP^{-1} \tag{33}$$

where D is a diagonal matrix made up of the eigenvalues:

$$D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \tag{34}$$

and the P's are matrices you make out of the eigenvalues.

This is both deep and useful; to deal with the useful aspect first, if

$$A = PDP^{-1} \tag{35}$$

then

$$A^2 = PDP^{-1}PDP^{-1} = PD^2P^{-1} \tag{36}$$

and, diagonal matrices are easy to deal with, if $D = \text{diag}(d_1, d_2, \ldots, d_n)$ then $D^2 = \text{diag}(d_1^2, d_2^2, \ldots, d_n^2)$. In this way diagonalizing a matrix can make lots of things you might want to do with the matrix easier.

The deep part is harder to see but it says in a way that for any matrix there is a set of axes where the matrix just scales along the axes. P^{-1} is a matrix, so it makes a linear transformation on your data, like a change of axes, so the equation is saying, first use P^{-1} to changes axes, then use D to scale the axes and then use P to go back to your original axes: every matrix has a system of axes where all it does is scale things!

In actual computational terms, we make P using the eigenvectors as columns, so $P = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. As we pointed out there is some question of the length of the eigenvectors, it is convenient in this context to make them unit length. For the example above then:

$$P = \sqrt{1}\sqrt{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{37}$$

and

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{\mathsf{T}} \tag{38}$$

A first look at PCA

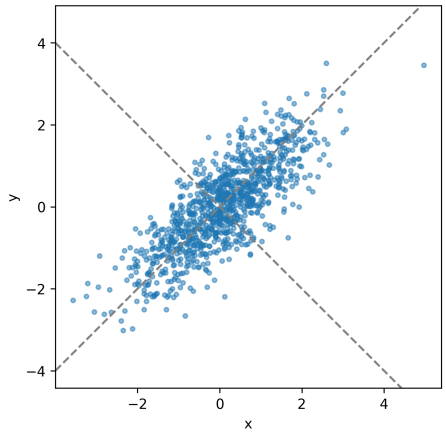

Figure 1: Anisotropic Gaussian with principal axes $(1, 1)$ and $(1, -1)$. You can come up with whatever story you like about these data, lets say that the ear size of lots of people has been measured, the x axis is the size of the left ear, the y axis the size of the right ear. For simplicity the average has been removed. Now, clearly the $x = y$ direction measures "how big is the person" and the $x = -y$ measures "how symmetric is the person".

In Figure 1 we have data points with a clear structure, they are spread out along the $x = y$ axis and have a bit of "blurring" along the $x = -y$ direction. In an experiment it might be that the $x = y$ represented the phenomena, and $x = -y$ noise, you might actually want to ignore the $x = -y$ direction and just project the points onto the $x = y$ direction; in any case, discovering this structure would be interesting. That's the business of PCA.

The first thing we might want to examine is the covariance matrix for the data:

$$\Sigma = \mathbb{E}\left[(x - \mu)(x - \mu)^{\mathsf{T}}\right]. \tag{39}$$

where μ is the mean and in our example we've already gotten rid of that. In

practice we work out the sample covariance:

$$\hat{\Sigma}_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}\big(x_i^{(k)} - \bar{x}_i\big)\big(x_j^{(k)} - \bar{x}_j\big). \tag{40}$$

The little hat is there to distinguish the theoretical concept with expectation values and what we actually work out from data.

For the data in Fig. 1 this give

$$\hat{\Sigma} = \begin{pmatrix} 1.20532775 & 0.92028907 \\ 0.92028907 & 1.12193513 \end{pmatrix} \tag{41}$$

and so the 1.20 is the variance in the x direction, 1.12 in the y direction and the 0.92 is how much does x depend on y. The data we measure, x and y is covariant, what we are interested in discovering, which in this simple example is clear from the graph, is the direction $x = y$ where the real action is happening; the data is something like "every data point has some value that determines how big x and y both are, except there is a little bit of noise that makes them a bit different. In this simple example the noise is completely unrelated to the signal, so knowing how far you are in the $x = y$ direction does not tell you anything about what is happening in the $x = -y$ direction. For real data this is not often true, but it is often approximately true. In any case, we can see that we are looking for directions where the covariance matrix is diagonal! This is diagonalization, what we looked at before.

If we work out the characteristic equation of Σ above we get $\lambda_1 = 2.07$ with eigenvector $\mathbf{e}_1 = (1, 1)^T$ and $\lambda_2 = 0.26$ with eigenvector $\mathbf{e}_2 = (1, -1)^T$, so, basically

$$\hat{\Sigma} = \begin{pmatrix} 1.20532775 & 0.92028907 \\ 0.92028907 & 1.12193513 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 2.07 & 0 \\ 0 & 0.26 \end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{\mathsf{T}} \tag{42}$$

So what's happened here: basically if we have already removed the average, the covariance matrix is

$$\Sigma_{\mathbf{x}} = \mathbb{E}\big[\mathbf{x}\mathbf{x}^{\mathsf{T}}\big]\,. \tag{43}$$

What happens if we do a change of variables, imagine $\mathbf{x} = P\mathbf{y}$

$$\begin{aligned} \Sigma_{\mathbf{x}} &= \mathbb{E}\big[\mathbf{x}\mathbf{x}^{\mathsf{T}}\big] \\ &= \mathbb{E}\big[P\mathbf{y}\mathbf{y}^{\mathsf{T}}P^{mathsfT}\big] = P\mathbb{E}\big[\mathbf{y}\mathbf{y}^{\mathsf{T}}\big]\,P^{\mathsf{T}} \\ &= P\Sigma_{\mathbf{y}}P^{\mathsf{T}} \end{aligned} \tag{44}$$

In otherwords, diagonalizing the covariance matrix is just a way to find new axes whose covariance matrix is diagonal. These new axes are the *principal components*.

The 2024 Irish Election

Elections to the Dáil, the Irish parliament uses the greatest voting system every created: single transferable vote. Clearly England uses the worst, first past the post; which is obviously terrible and not worth discussing. The advantage of STV over the sort of list systems employed in Europe is more subtle; a list system tries to match the number of representative to the proportion of people voting for them, if 25% of people support party A then party A should have roughly 25% of the seats. STV does something better, it represents the division inside the hearts of individuals, rather than the division across the electorate; it is a beautiful idea, the make up of the Dáil represents not just how views change from person to person, but the conflicting views inside each individual. I urge you to look at how STV works.

We are going to ignore that for now though and look at only the first preference vote, each person marks with a one the person they would mostly like to see elected in their constituency. Ireland has 43 constituencies and elects 174 representatives, called TDs or Teachtaí Dála, to the Dáil. There are a lot of parties, the two historic parties of power Fianna Fáil and Fine Gael, the insurgent Sinn Féin, actually the oldest party in the state but never in power since it lost the Irish Civil War in 1922, a few left wing parties, such as the Labour Party, the Social Democrats and People before Profit, along with some right wing parties such as Aontú and Independent Ireland; many constituencies also have independent candidates who focus on local issues, for example politics in Kerry is dominated by a single family, the Healy-Raes who have few principles other than always wearing cloth caps and trying to get money for Kerry.

There was an election in 2024; it saw the Green Party lose most of its seats since it had annoyed its supporters by joining a coalition government; Sinn Féin didn't do as well as had been expected and a joint government formed of Fianna Fáil and Fine Gael remained in power, this coalition included some independents and dropped the Green Party. All this means that the data for this election includes 43 data points, one for each constituency and each data point is eleven dimensional, corresponding to voter share for 10 political

parties and a final category of fringe parties and independents. Not every party runs candidates in every constituency, for convenience the blank entries are replaced with zeros.

If you do PCA on these data you end up with 11 eigenvalues, strongly dominated by the first three or four, after that the get small. This indicates that there are probably two or three main factors that determine how people vote. By doing PCA, in this case using the correlation matrix rather than the covariance data, this is a technical detail made necessary by the fact the numbers always add up to one. Fig. 2 shows the result for the first and second component coloured by the third.

Summary

Glossary

1 and **0** these are an 'abused notation', I am relying on your common sense to guess what they mean in context, **1** might mean

$$\mathbf{1} = \mathrm{diag}(1, 1, \ldots, 1) \tag{45}$$

or it might be a vector with just ones, similarly **0** might be a matrix with only zeros or it might be vector with only zeros. Mathematics, the supposed exactest of sciences, is often filled with 'abused notation'; it is an interesting development that mathematicians these days are creating a computer language approach called `Leen` that allows you, or forces you, to write down mathematics in an exact way, which is useful and good for mathematics, but not for communicating mathematics where stopping all the time to give definitions can make things very dry and boring to read.

The symbol δ_{ij} is the *Kronecker delta*:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{46}$$

The *covariance matrix*, if $\mathbf{X} = (X_1, X_2, \ldots, X_n)^\mathsf{T}$ is a vector of random variables

$$\Sigma = \mathrm{Cov}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top\right], \qquad \mu_i = \mathbb{E}[X_i]. \tag{47}$$

or, component-wise

$$\Sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]\,. \tag{48}$$

The *correlation matrix* is a *whitened* version of this:

$$\sigma_i^2 = \mathrm{Var}(X_i) = \Sigma_{ii}, \qquad \sigma_i = \sqrt{\Sigma_{ii}}. \tag{49}$$

and let

$$D = \mathrm{diag}(\sigma_1, \ldots, \sigma_n). \tag{50}$$

then

$$R = D^{-1}\Sigma D^{-1}. \tag{51}$$

or component-wise

$$R_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sigma_i\,\sigma_j} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}. \tag{52}$$

When you are doing PCA you should use the covariance if that units are the same for all the dimensions and the data variance is somehow comparable, and correlation otherwise. The election data is a sort of edge case, the correlation matrix produces better results because some of the parties have a much bigger vote than others, there is also a technical complication related to the entries in the data vector always adding up to one.

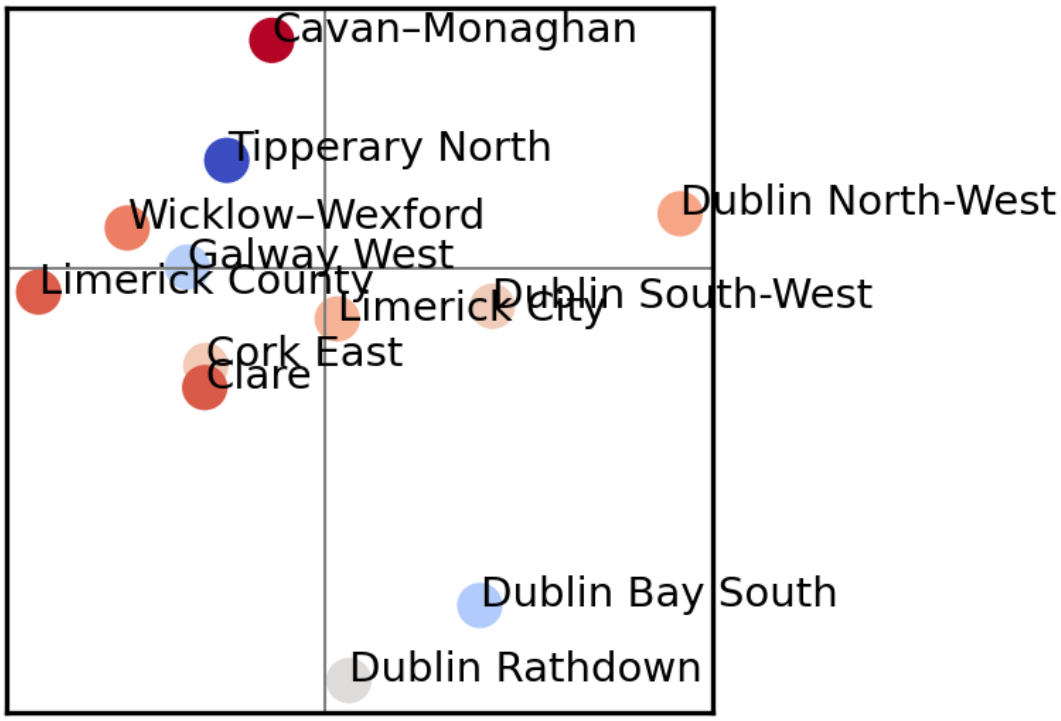

Figure 2: This shows the first three principal components, the first in the x-axis, the second, the y-axis and the third by colour. Only 12 data points, picked randomly, are shown otherwise it is hard to read. The y-axis is clearly a left-right axis, Dublin-Rathdown is a very liberal place, Cavan-Monaghan very conservative; the x-axis is not as clear but seems to measure diversity of opinion, the lefter points correspond to more complex places mixing rural and urban areas, for example. The third component, illustrated by colour, is interesting, it seems to correspond to independent candidates doing well, the two blues places, Galway West and Tipperary North have strong local candidates, Seamus Healy in Tipperary North and Catherine Connolly, who subsequently ran successfully for the presidency, in Galway West.