Optimization Models

EECS 127 / EECS 227AT

Laurent El Ghaoui

EECS department
UC Berkeley

Fall 2018

LECTURE 5

Symmetric Matrices

Whoever think algebra is a trick in obtaining unknowns has thought in vain. No attention should be paid to the fact that algebra and geometry are different in appearance. Algebras are geometric facts which are proved.

Omar Khayyam, 1050–1123.

Outline

1. Basics

2. The spectral theorem
   - Variational characterization of eigenvalues

3. Positive-semidefinite matrices
   - Matrix square-root and Cholesky decomposition
   - Ellipsoids
   - The PSD cone and partial order
   - Schur complements

4. Principal Component Analysis
   - Overview
   - Iterated variance maximization
   - Example

Basics

- A square matrix $A \in \mathbb{R}^{n,n}$ is *symmetric* if it is equal to its transpose: $A = A^\top$, that is: $A_{ij} = A_{ji}$, $1 \leq i, j \leq n$.
- Elements above the diagonal in a symmetric matrix are thus identical to corresponding elements below the diagonal.
- Symmetric matrices are ubiquitous in engineering applications. They arise, for instance, in the description of graphs with undirected weighted edges between the nodes, in geometric distance arrays (between, say, cities), in defining the Hessian of a nonlinear function, in describing the covariances of random vectors, etc.
- The following is an example of a 3 × 3 symmetric matrix:

$$A = \begin{bmatrix} 4 & 3/2 & 2 \\ 3/2 & 2 & 5/2 \\ 2 & 5/2 & 2 \end{bmatrix}.$$

- The set of symmetric $n \times n$ matrices is is a subspace of $\mathbb{R}^{n,n}$, and it is denoted with $\mathbb{S}^n$.

Example

Example 1 (Sample covariance matrix)

- *Given m points $x^{(1)}, \dots, x^{(m)}$ in $\mathbb{R}^n$, we define the sample covariance matrix to be the $n \times n$ symmetric matrix*

$$C \doteq \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \hat{x})(x^{(i)} - \hat{x})^\top,$$

where $\hat{x} \in \mathbb{R}^n$ is the sample average of the points: $\hat{x} \doteq \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$.

- *The covariance matrix C is obviously a symmetric matrix. This matrix arises when computing the sample variance of the scalar products $s_i \doteq w^\top x^{(i)}$, $i = 1, \dots, m$, where $w \in \mathbb{R}^n$ is a given vector:*

$$\sigma^2 = \sum_{i=1}^{m} (w^\top x^{(i)} - \hat{s})^2 = \sum_{i=1}^{m} (w^\top (x^{(i)} - \hat{x}))^2 = w^\top C w.$$

Example

Example 2 (Portfolio variance)

- *For n financial assets, we can define a vector $r \in \mathbb{R}^n$ whose components r_k are the rate of returns of the k-th asset, $k = 1, \dots, n$.*
- *Assume now that we have observed m samples of historical returns $r^{(i)}$, $i = 1, \dots, m$. The sample average over that history of return is $\hat{r} = (1/m)(r^{(1)} + \dots + r^{(m)})$, and the sample covariance matrix has (i, j) component given by*

$$C_{ij} = \frac{1}{m} \sum_{t=1}^{m} (r_i^{(t)} - \hat{r}_i)(r_j^{(t)} - \hat{r}_j), \ \ 1 \leq i, \ \ j \leq n.$$

- *If $w \in \mathbb{R}^n$ represents a portfolio "mix," that is $w_k \geq 0$ is the fraction of the total wealth invested in asset k, then the return of such a portfolio is given by $\rho = r^\top w$.*
- *The sample average of the portfolio return is $\hat{r}^\top w$, while the sample variance is given by $w^\top C w$.*

Basics

Example 3 (Hessian matrix)

- *The Hessian of a twice differentiable function* $f : \mathbb{R}^n \to \mathbb{R}$ *at a point* $x \in \operatorname{dom} f$ *is the matrix containing the second derivatives of the function at that point. That is, the Hessian is the matrix with elements given by*

$$H_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \;\; 1 \leq i, \;\; j \leq n.$$

- *The Hessian of* f *at* x *is often denoted as* $\nabla^2 f(x)$.
- *Since the second-derivative is independent of the order in which derivatives are taken, it follows that* $H_{ij} = H_{ji}$ *for every pair* (i, j)*, thus the Hessian is always a symmetric matrix.*

Basics

Quadratic functions

- Consider the quadratic function (a polynomial function is said to be quadratic if the maximum degree of its monomials is equal to two)

$$q(x) = x_1^2 + 2x_1x_2 + 3x_2^2 + 4x_1 + 5x_2 + 6.$$

- The Hessian of q at x is given by

$$H = \left[\frac{\partial^2 q(x)}{\partial x_i \partial x_j}\right]_{1 \le i,j \le 2} = \left[\begin{array}{cc} \frac{\partial q}{\partial x_1^2} & \frac{\partial^2 q}{\partial x_1 \partial x_2} \\ \frac{\partial^2 q}{\partial x_2 \partial x_1} & \frac{\partial q}{\partial x_2^2} \end{array}\right] = \left[\begin{array}{cc} 2 & 2 \\ 2 & 6 \end{array}\right].$$

- The monomials in $q(x)$ of degree two can also be written compactly as

$$x_1^2 + 2x_1x_2 + 3x_2^2 = \frac{1}{2}x^\top Hx.$$

- Any quadratic function can be written as the sum of a quadratic term involving the Hessian, and an affine term:

$$q(x) = \frac{1}{2}x^\top Hx + c^\top x + d, \quad c^\top = [4 \;\; 5], \; d = 6.$$

The spectral theorem

Any symmetric matrix is orthogonally similar to a diagonal matrix. This is stated in the following so-called *spectral theorem* for symmetric matrices.

Theorem 1 (Spectral Theorem)

Let $A \in \mathbb{R}^{n,n}$ *be symmetric, let* $\lambda_i \in \mathbb{R}$, $i = 1, \ldots, n$, *be the eigenvalues of* A *(counting multiplicities). Then, there exist a set of orthonormal vectors* $u_i \in \mathbb{R}^n$, $i = 1, \ldots, n$, *such that* $Au_i = \lambda_i u_i$. *Equivalently, there exist an orthogonal matrix* $U = [u_1 \cdots u_n]$ *(i.e.,* $UU^\top = U^\top U = I_n$*) such that*

$$A = U \Lambda U^\top = \sum_{i=1}^{n} \lambda_i u_i u_i^\top, \quad \Lambda = \mathrm{diag}\left(\lambda_1, \ldots, \lambda_n\right).$$

Variational characterization of eigenvalues

- Since the eigenvalues of $A \in \mathbb{S}^n$ are real, we can arrange them in decreasing order:

$$\lambda_{\max}(A) = \lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A) = \lambda_{\min}(A).$$

- The extreme eigenvalues can be related to the minimum and the maximum attained by the quadratic form induced by A over the unit Euclidean sphere.
- For $x \neq 0$ the ratio

$$\frac{x^\top A x}{x^\top x}$$

  is called a *Rayleigh quotient*.

Variational characterization of eigenvalues

Theorem 2 (Rayleigh quotients)

Given $A \in \mathbb{S}^n$, it holds that

$$\lambda_{\min}(A) \leq \frac{x^\top A x}{x^\top x} \leq \lambda_{\max}(A), \quad \forall x \neq 0.$$

Moreover,

$$\begin{aligned} \lambda_{\max}(A) &= \max_{x:\, \|x\|_2 = 1} x^\top A x \\ \lambda_{\min}(A) &= \min_{x:\, \|x\|_2 = 1} x^\top A x, \end{aligned}$$

and the maximum and minimum are attained for $x = u_1$ and for $x = u_n$, respectively, where u_1 (resp. u_n) is the unit-norm eigenvector of A associated with its largest (resp. smallest) eigenvalue of A.

Matrix gain

- Given a matrix $A \in \mathbb{R}^{m,n}$, let us consider the linear function associated to A, which maps input vectors $x \in \mathbb{R}^n$ to output vectors $y \in \mathbb{R}^m$:

$$y = Ax.$$

- Given a vector norm, the matrix *gain*, or operator norm, is defined as the maximum value of the ratio $\|Ax\|/\|x\|$ between the size (norm) of the output and the of the input.
- In particular, the gain with respect to the Euclidean norm is defined as

$$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2},$$

  and it is often referred to as the *spectral* norm of A.
- The square of the input-output ratio in the Euclidean norm is

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} = \frac{x^\top (A^\top A) x}{x^\top x}$$

Matrix gain

- This quantity is upper and lower bounded by the maximum and by the minimum eigenvalue of the symmetric matrix $A^\top A \in \mathbb{S}^n$, respectively:

$$\lambda_{\min}(A^\top A) \leq \frac{\|Ax\|_2^2}{\|x\|_2^2} \leq \lambda_{\max}(A^\top A)$$

- The upper and lower bounds are actually attained when x is equal to an eigenvector of $A^\top A$ corresponding respectively to the maximum and to minimum eigenvalue of $A^\top A$. Therefore,

$$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\lambda_{\max}(A^\top A)},$$

  where this maximum gain is obtained for x along the direction of eigenvector u_1 of $A^\top A$, and

$$\min_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\lambda_{\min}(A^\top A)},$$

  where this minimum gain is obtained for x along the direction of eigenvector u_n of $A^\top A$.

Positive-semidefinite matrices

- A symmetric matrix $A \in \mathbb{S}^n$ is said to be *positive semidefinite* (PSD) if the associated quadratic form is nonegative, i.e.,

$$x^\top Ax \geq 0, \quad \forall x \in \mathbb{R}^n.$$

- If, moreover,

$$x^\top Ax > 0, \quad \forall 0 \neq x \in \mathbb{R}^n,$$

  then A is said to be *positive definite*. To denote a symmetric positive semidefinite (resp. positive definite) matrix, we use the notation $A \succeq 0$ (resp. $A \succ 0$).

- We say that A is negative semidefinite, written $A \preceq 0$, if $-A \succeq 0$, and likewise A is negative definite, written $A \prec 0$, if $-A \succ 0$.

- It is immediate to see that a positive semidefinite matrix is actually positive definite if and only if it is invertible.

- It holds that

$$\begin{aligned} A \succeq 0 \quad &\Leftrightarrow \quad \lambda_i(A) \geq 0, \; i = 1, \ldots, n \\ A \succ 0 \quad &\Leftrightarrow \quad \lambda_i(A) > 0, \; i = 1, \ldots, n. \end{aligned}$$

Congruence transformations

Corollary 1

For any matrix $A \in \mathbb{R}^{m,n}$ *it holds that:*

1. $A^\top A \succeq 0$, *and* $AA^\top \succeq 0$;
2. $A^\top A \succ 0$ *if and only if* A *is full-column rank, i.e.,* $\operatorname{rank} A = n$;
3. $AA^\top \succ 0$ *if and only if* A *is full-row rank, i.e.,* $\operatorname{rank} A = m$.

Matrix square-root and Cholesky decomposition

- Let $A \in \mathbb{S}^n$. Then

$$\begin{aligned} A \succeq 0 \quad &\Leftrightarrow \quad \exists\, B \succeq 0 : \; A = B^2 \\ A \succ 0 \quad &\Leftrightarrow \quad \exists\, B \succ 0 : \; A = B^2. \end{aligned}$$

- Matrix $B = A^{1/2}$ is called the *matrix square-root* of A.
- Any $A \succeq 0$ admits the spectral factorization $A = U \Lambda U^\top$, with U orthogonal and $\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)$, $\lambda_i \geq 0$, $i = 1, \ldots, n$. Defining $\Lambda^{1/2} = \mathrm{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_1}\right)$ and $B = U \Lambda^{1/2} U^\top$:

$$\begin{aligned} A \succeq 0 \quad &\Leftrightarrow \quad \exists\, B : \; A = B^\top B \\ A \succ 0 \quad &\Leftrightarrow \quad \exists\, B \text{ nonsingular} : \; A = B^\top B. \end{aligned}$$

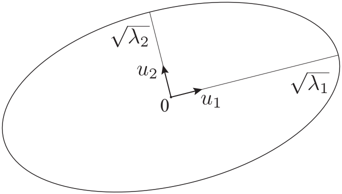
- A is positive definite if and only if it is congruent to the identity.

Positive-definite matrices and ellipsoids

- Positive-definite matrices are intimately related to geometrical objects called *ellipsoids*.
- A full-dimensional, bounded ellipsoid with center in the origin can indeed be defined as the set

$$\mathcal{E} = \{x \in \mathbb{R}^n : \ x^\top P^{-1} x \leq 1\}, \quad P \succ 0.$$

- The eigenvalues λ_i and eigenvectors u_i of P define the orientation and shape of the ellipsoid: u_i the directions of the semi axes of the ellipsoid, while their lengths are given by $\sqrt{\lambda_i}$.
- Using the Cholesky decomposition $P^{-1} = A^\top A$, the previous definition of ellipsoid $\mathcal{E}$ is also equivalent to: $\mathcal{E} = \{x \in \mathbb{R}^n : \ \|Ax\|_2 \leq 1\}$.



The PSD cone and partial order

- The set of positive semidefinite matrices $\mathbb{S}^n_+$ is a *convex cone*.
- First, it is a *convex* set, since it satisfies the defining property of convex sets (more on this later!), that is for any two matrices $A_1, A_2 \in \mathbb{S}^n_+$ and any $\theta \in [0, 1]$, it holds that
$$x^\top(\theta A_1 + (1-\theta)A_2)x = \theta x^\top A_1 x + (1-\theta)x^\top A_2 x \geq 0, \ \forall x,$$
hence $\theta A_1 + (1-\theta)A_2 \in \mathbb{S}^n_+$.
- Moreover, for any $A \succeq 0$ and any $\alpha \geq 0$, we have that $\alpha A \succeq 0$, which says that $\mathbb{S}^n_+$ is a *cone*.
- The relation "$\succeq$" defines a partial order on the cone of PSD matrices. That is, we say that $A \succeq B$ if $A - B \succeq 0$ and, similarly, $A \succ B$ if $A - B \succ 0$. This is a *partial* order, since not any two symmetric matrices may be put in a $\preceq$ or $\succeq$ relation.

Schur complements

Theorem 3 (Schur complements)

Let $A \in \mathbb{S}^n$, $B \in \mathbb{S}^m$, $X \in \mathbb{R}^{n,m}$, with $B \succ 0$. Consider the symmetric block matrix

$$M = \begin{bmatrix} A & X \\ X^\top & B \end{bmatrix},$$

and define the so-called Schur complement matrix of A in M

$$S \doteq A - XB^{-1}X^\top.$$

Then,

$$M \succeq 0 \text{ (resp., } M \succ 0) \quad \Leftrightarrow \quad S \succeq 0 \text{ (resp., } S \succ 0).$$

Principal Component Analysis

Motivation

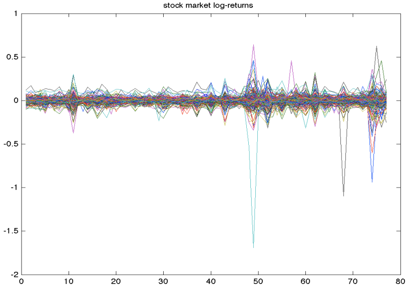


Figure: Daily log-returns of 77 Fortune 500 companies, 1/2/2007—12/31/2008.

High-dimensional data does not make any sense! (Other than tell us: returns are approximately zero . . .)

Which view is better?

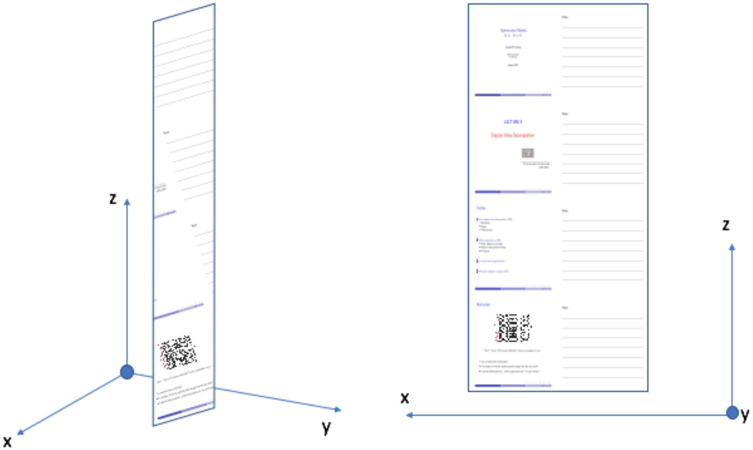


Figure: A “flat” data set viewed from two different angles in $\mathbb{R}^3$.

Principal Component Analysis

Overview

Principal Component Analysis (PCA) originated in psychometrics in the 1930's. It is now widely used in

- Exploratory data analysis.
- Simulation.
- Visualization.

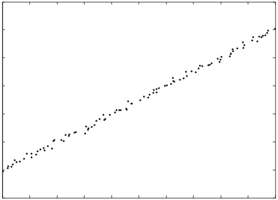
Application fields include

- Finance, marketing, economics.
- Biology, medecine.
- Engineering design, signal compression and image processing.
- Search engines, data mining.

Principal component analysis (PCA)

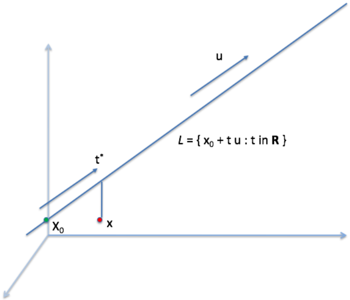
Basic idea

- Principal component analysis (PCA) is a technique of *unsupervised learning*, widely used to "discover" the most important, or informative, directions in a data set, that is the directions along which the data varies the most.
- In the data cloud below it is apparent that there exist a direction (at about 45 degrees from the horizontal axis) along which almost all the variation of the data is contained. In contrast, the direction at about 135 degrees contains very little variation of the data.



- The important direction was easy to spot in this two-dimensional example. However, graphical intuition does not help when analyzing data in dimension $n > 3$, which is where Principal Components Analysis (PCA) comes in handy.

Recap from lecture 2: projection on a line



A line is an affine subspace of dimension 1. It can be parametrized as

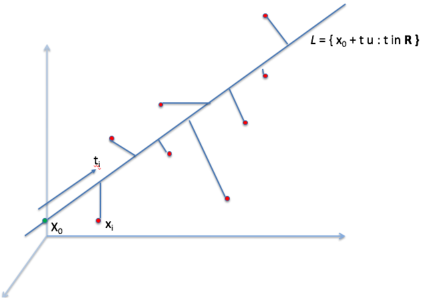
$$\mathcal{L} = \{x_0 + tu \; : \; t \in \mathbb{R}\},$$

where $x_0, u \in \mathbb{R}^n$ are given, with $\|u\|_2 = 1$.

Projection of a point $x \in \mathbb{R}$ on the line:

$$z^* = \arg\min_{z \in \mathcal{L}} \; \|z - z\|_2 = x_0 + t^* u, \quad t^* = u^\top (x - x_0).$$

Variance of scores of projected points



Given a line ($x_0, u \in \mathbb{R}^n$ are given, with $\|u\|_2 = 1$)

$$\mathcal{L} = \{x_0 + tu \ : \ t \in \mathbb{R}\},$$

we seek to find the variance of the scores $t_i = u^T(x_i - x_0)$, $i = 1, \ldots, m$ of the projected points.

This variance of the scores is

$$(1/m)\sum_{i=1}(u^\top x_i - u^\top \hat{x})^2 = u^\top C u,$$

where $\hat{x} = (1/m)(x_1 + \ldots + x_m)$, and C is the covariance matrix of the data.

Variance maximization problem

Let C be the (empirical) covariance matrix. *Variance maximization problem:*

$$\max_x \ u^\top Cu \ : \ \|u\|_2 = 1.$$

Assume the EVD of C is given:

$$C = \sum_{i=1}^{p} \lambda_i u_i u_i^\top ,$$

with $\lambda_1 \geq \ldots \lambda_p$, and $U = [u_1, \ldots, u_p]$ is orthogonal $(U^\top U = I)$. Then a solution to

$$\max_{u \,:\, \|u\|_2 = 1} \ u^\top Cu$$

is $u^* = u_1$, with u_1 an eigenvector of C that corresponds to its largest eigenvalue λ_1.

Alternatively, u_1 can be found directly via SVD of (centered) data matrix (see later).

Finding orthogonal directions

A deflation method

Once we've found a direction with high variance, can we repeat the process and find other ones?

Deflation method:

- Project data points on the subspace orthogonal to the direction we found.
- Find a direction of maximal variance for projected data.

The process stops after p steps (p is the dimension of the whole space), but can be stopped earlier (to find only k directions, with $k << p$).

Finding orthogonal directions

Result

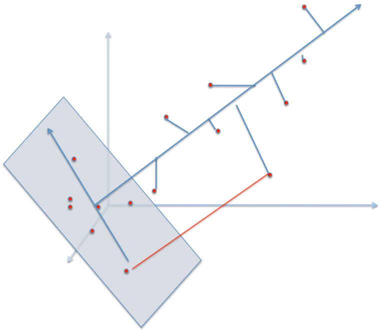
It turns out that the direction that solves

$$\max_u \ u^\top C u \ : \ u^\top u_1 = 0, \ \ \|u\|_2 = 1,$$

is u_2, an eigenvector corresponding to the second-to-largest eigenvalue.

After k steps of the deflation process, the directions returned are $u_1, \ldots, u_k$. Thus we can compute k directions of largest variance in *one* eigenvalue decomposition of the covariance matrix.

Geometry of deflation



In PCA, we first identify a line $\mathcal{L}$ such that the points projected on $\mathcal{L}$ have high variance.

Deflation consists in projecting the data on a hyperplane orthogonal to the line $\mathcal{L}$; a new minimum-distance line contained in the hyperplane is then found.

Since the dimension of the problem is reduced by one at each step, this process stops in at most n steps.

Measuring quality

How well is data approximated by its projections on the successive subspaces?

Approach: compare sum of variances contained in the k directions found, with total variance.

Explained variance: measured by the ratio

$$\frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_p},$$

where $\lambda_1 \geq \ldots \geq \lambda_p$ are the eigenvalues of the covariance matrix.

Examples

PCA of market data, prior to 2008 crisis

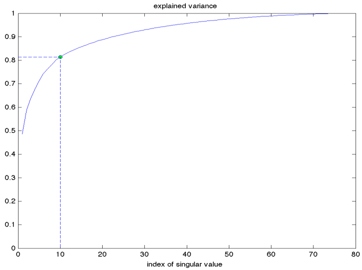


Figure: Data: Daily log-returns of 77 Fortune 500 companies, 1/2/2007—12/31/2008.

- Plot shows the eigenvalues of covariance matrix in decreasing order.
- First ten components explain 80% of the variance.
- Largest magnitude of eigenvector for 1st component correspond to financial sector (FABC, FTU, MER, AIG, MS).

Examples

PCA of voting data

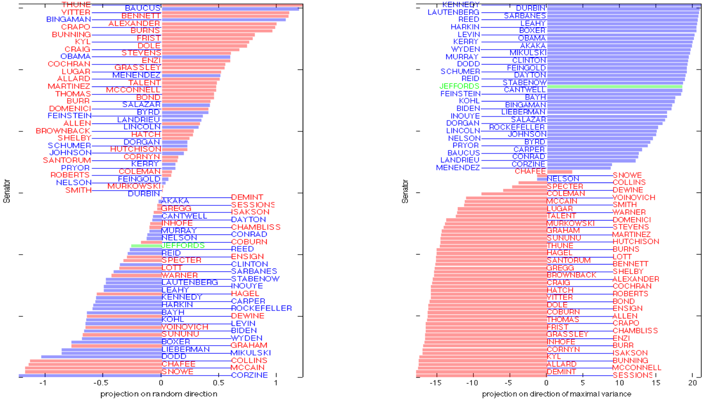


Figure: Projection of US Senate voting data on random direction (left panel) and direction of maximal variance (right panel). The latter reveals party structure (party affiliations added after the fact). Note also the much higher range of values it provides.