Axis Calibration for Improving Data Attribute Estimation in Star Coordinates Plots: Supplemental Material

Manuel Rubio-Sánchez and Alberto Sanchez

1 Notation

Table 1 summarizes the most relevant notation used throughout the paper.

2 User estimates on individual data sets

Figure 1 shows average estimation errors $\upsilon(n)$ obtained in experiments with users for SC, and the automatic estimation onto calibrated axes, on (a) the breakfast cereal, and (b) wine data sets.

3 Model SC_M as a linear program

In practice, users estimate data values in SC through the following convex optimization problem (denoted as SC_M):

$$\begin{array}{ll} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} & \|\mathbf{D}\mathbf{x}\|_1 \\ \text{subject to} & \mathbf{V}^\mathsf{T}\mathbf{x}=\mathbf{p}, \\ & \mathbf{0}\preceq\mathbf{x}\preceq\mathbf{1}, \end{array}$$

where $\mathbf{D}$ is diagonal matrix where $d_{i,i}=\|\mathbf{v}_i\|_2$. Note that $\|\mathbf{D}\mathbf{x}\|_1$ corresponds to the length of the path. Alternatively, it can be rewritten as a linear program (LP):

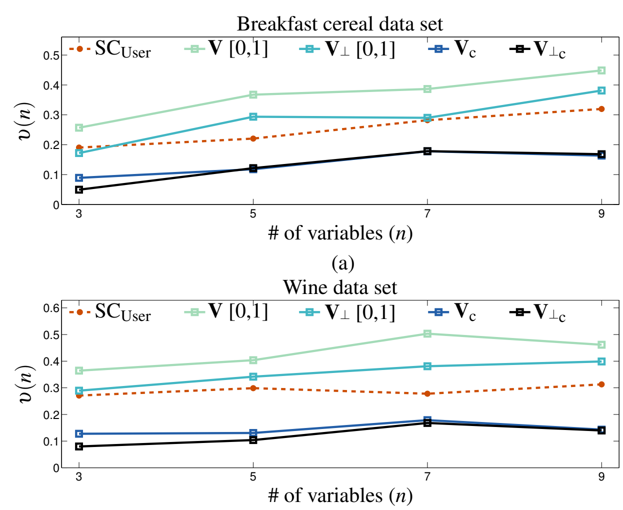
$$\begin{array}{ll} \underset{\mathbf{t}\in\mathbb{R}^n,\ \mathbf{x}\in\mathbb{R}^n}{\text{minimize}} & \mathbf{1}^\mathsf{T}\mathbf{t} \\ \text{subject to} & -\mathbf{t}\preceq\mathbf{D}\mathbf{x}\preceq\mathbf{t}, \\ & \mathbf{V}^\mathsf{T}\mathbf{x}=\mathbf{p}, \\ & \mathbf{0}\preceq\mathbf{x}\preceq\mathbf{1}. \end{array}$$

4 Sparse solutions for SC_M

Figure 2 shows distributions associated with the number of nonzero elements of the solutions provided by SC_M. The results include estimates on the four data sets used, where the number of variables is a multiple of 3, over 30 random vector configurations according to uniform distributions ($\mathscr{U}[0,2\pi]$ for angles, and $\mathscr{U}[0.5,1]$ for lengths), and subsets of variables also chosen at random. It is apparent that most solutions involve only two or three nonzero values, in accordance with the users' main strategy.

Table 1. Notation Summary.

n	Dimensionality of the data space
m	Dimensionality of the observable space (2 in this paper)
$\mathbf{V}$	$n\times m$ matrix of axis vectors
$\mathbf{v}_i$	i-th axis vector (i-th row of $\mathbf{V}$)
$\tilde{\mathbf{v}}_i$	i-th column vector of $\mathbf{V}$
$\mathbf{x}$	Data instance of dimensionality n
x_i	i-th data attribute of instance $\mathbf{x}$
$\hat{\mathbf{x}}$	Vector of estimates of $\mathbf{x}$
$\hat{x}_i$	estimate of x_i
$\mathbf{p}$	Low m-dimensional representation of a data instance
$\|\cdot\|, \|\cdot\|_2$	Euclidean norm
$\|\cdot\|_1$	ℓ_1 norm
$\|\cdot\|_{\text{Fro}}$	Matrix Frobenius norm
$\langle\cdot,\cdot\rangle$	Dot product
$\mathbf{0}$	Vector of all zeros
$\mathbf{1}$	Vector of all ones
$\mathbf{e}_i$	Vector of all zeros, except a 1 for the i-th component
$\mathbf{I}$	Identity matrix
ε	Estimation error for a data sample
$\mathscr{R}(\mathbf{V})$	Range (i.e., column space) of $\mathbf{V}$
$\mathbf{V}^\dagger$	Moore-Penrose pseudoinverse of matrix $\mathbf{V}$
$\mathbf{V}_\perp$	Orthogonal matrix for which $\mathscr{R}(\mathbf{V}_\perp)=\mathscr{R}(\mathbf{V})$, i.e., the result of orthonormalizing the columns of $\mathbf{V}$
$\mathscr{U}$	Uniform distribution
$\preceq$	Vector componentwise inequality
$\upsilon(n)$	Average estimation errors in user experiments
$\delta(n)$	Average estimation errors in automatic simulations



(b)

Fig. 1. Average estimation errors $\upsilon(n)$ on (a) the breakfast cereal, and (b) wine data sets.

- *Manuel Rubio-Sánchez is with URJC. E-mail: manuel.rubio@urjc.es.*
- *Alberto Sanchez is with URJC. E-mail: alberto.sanchez@urjc.es.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

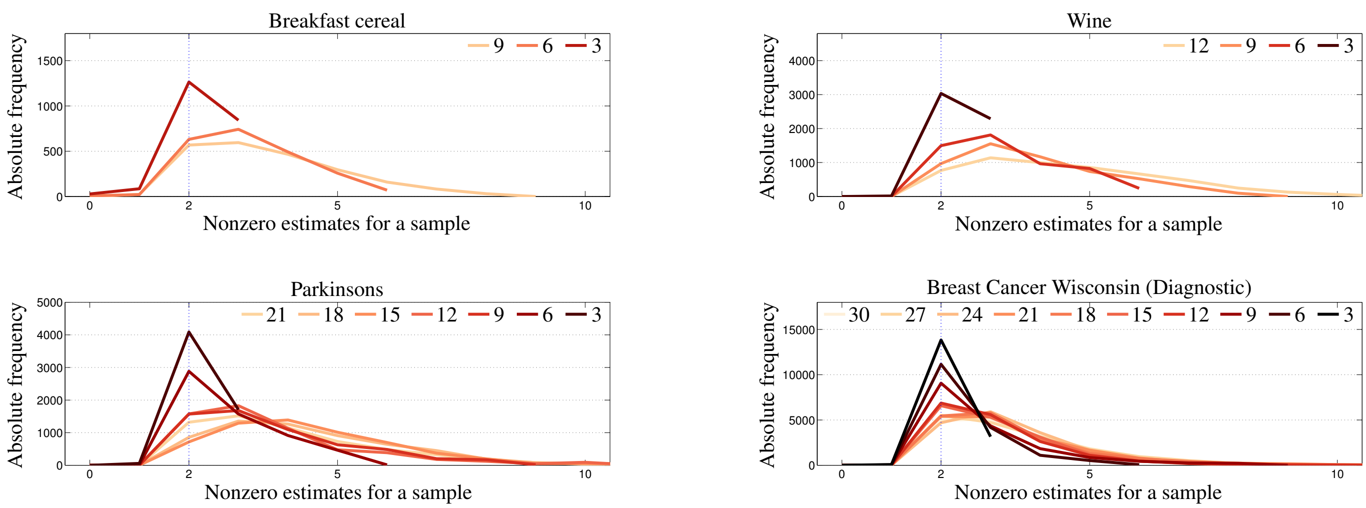


Fig. 2. Distributions associated with the number of nonzero elements of the solutions provided by SC_M