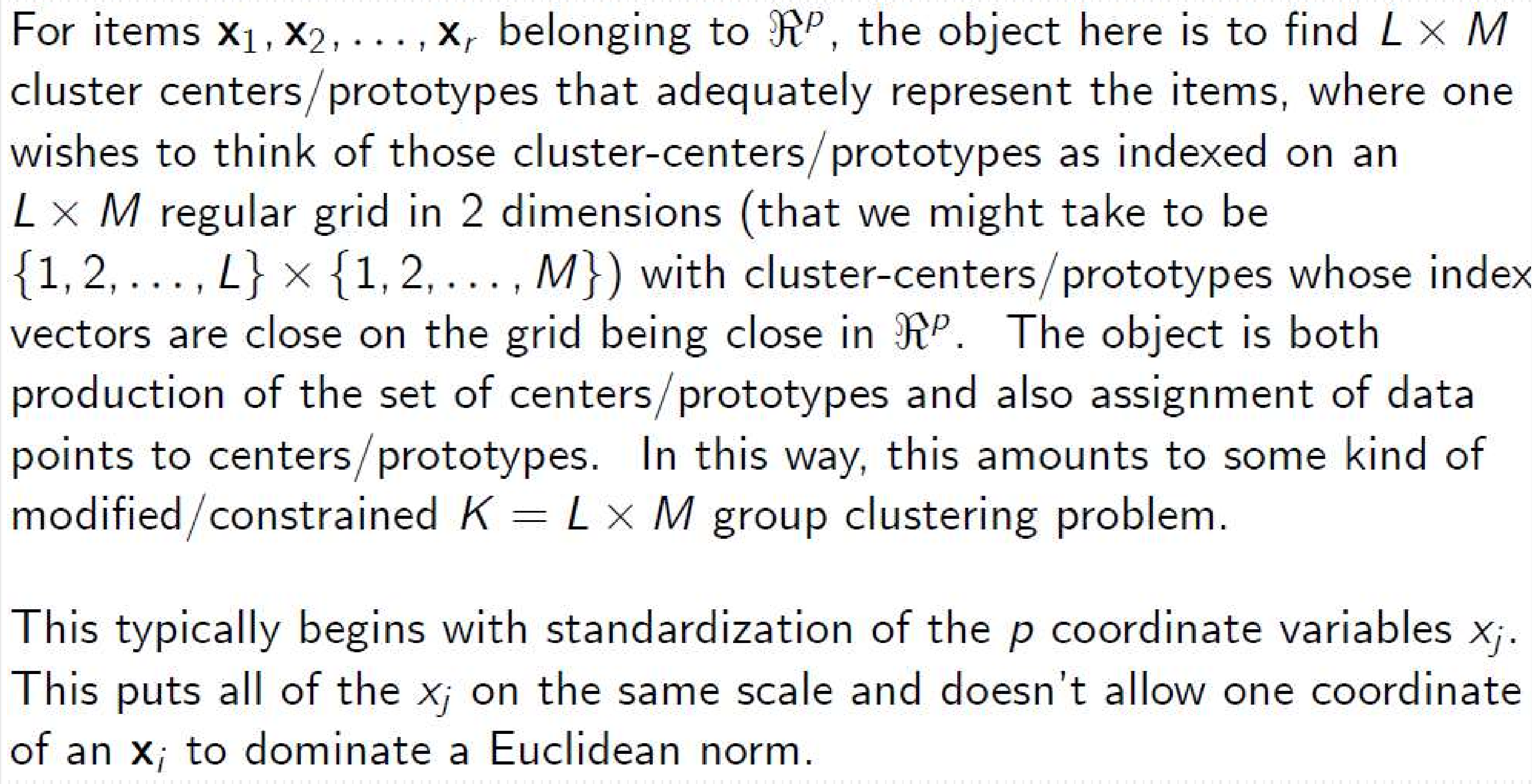
Self-Organizing Maps

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Objectives/preliminaries

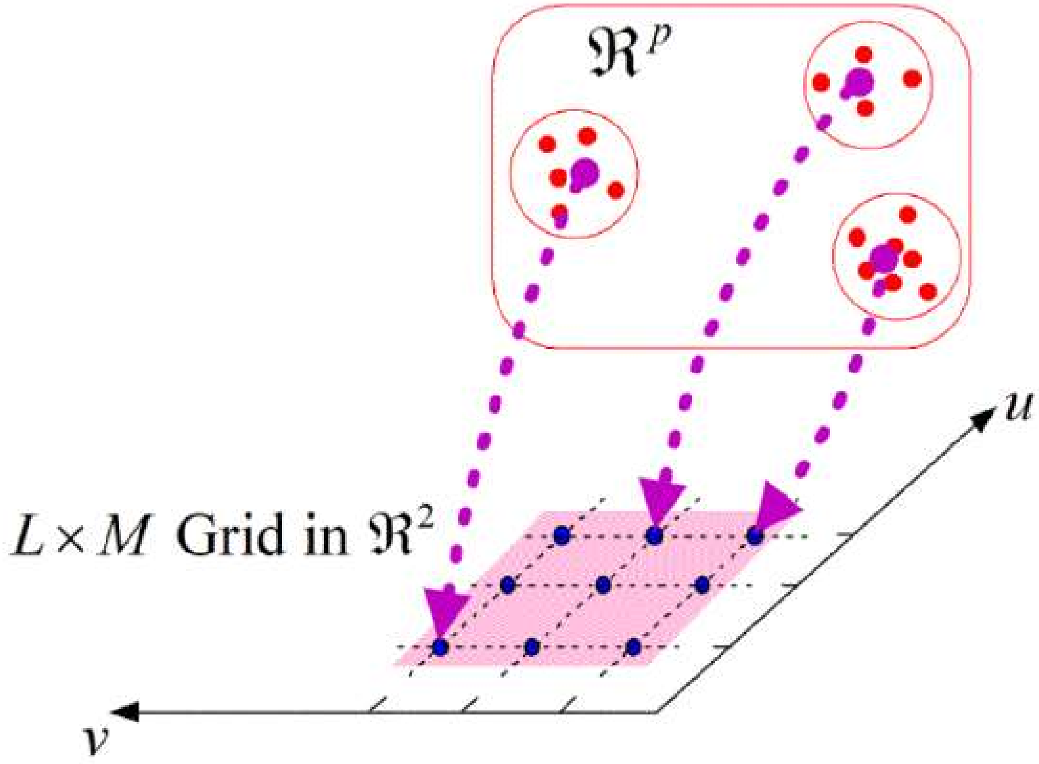


For items $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r$ belonging to $\Re^p$, the object here is to find $L \times M$ cluster centers/prototypes that adequately represent the items, where one wishes to think of those cluster-centers/prototypes as indexed on an $L \times M$ regular grid in 2 dimensions (that we might take to be $\{1, 2, \ldots, L\} \times \{1, 2, \ldots, M\}$) with cluster-centers/prototypes whose index vectors are close on the grid being close in $\Re^p$. The object is both production of the set of centers/prototypes and also assignment of data points to centers/prototypes. In this way, this amounts to some kind of modified/constrained $K = L \times M$ group clustering problem.

This typically begins with standardization of the p coordinate variables x_j. This puts all of the x_j on the same scale and doesn't allow one coordinate of an $\mathbf{x}_i$ to dominate a Euclidean norm.

Objectives/preliminaries cont.

Below is a cartoon illustrating the objective of making a SOM.

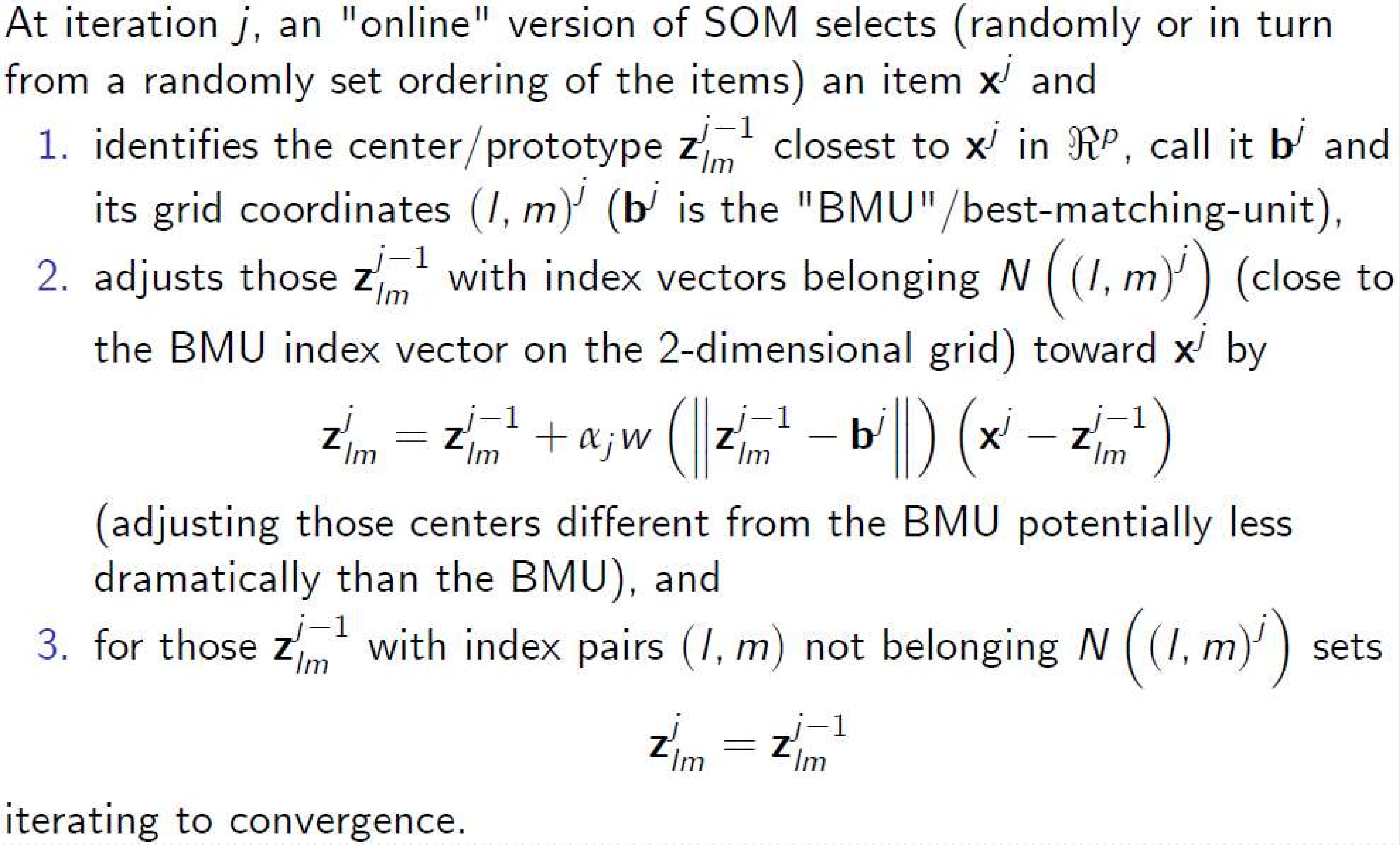


Kohonen’s SOM algorithms

Begin with some initial cluster centers $\left\{\mathbf{z}_{lm}^{0}\right\}_{l=1,\ldots,L \text{ and } m=1,\ldots,M}$ in $\Re^{p}$. This might be a random selection (without replacement or the possibility of duplication) from the set of items. It might be a set of grid points in the 2-d "plane" in $\Re^{p}$ defined by the first two principal components of the items $\{\mathbf{x}_i\}_{i=1,\ldots,r}$.

Then define neighborhoods on the $L \times M$ grid, $N(l,m)$ (in $\Re^{2}$), that are subsets of the grid "close" to the various elements of the $L \times M$ grid. $N(l,m)$ could be all of the grid, (l,m) alone, all grid points (l',m') within some constant 2-dimensional Euclidean distance, of (l,m), etc. Define a weighting function (for $\Re^{p}$), say $w(\|\mathbf{x}\|)$, so that $w(0)=1$ and $w(\|\mathbf{x}\|) \geq 0$ is monotone non-increasing in $\|\mathbf{x}\|$. For some schedule of non-increasing positive constants $1 > \alpha_1 \geq \alpha_2 \geq \alpha_3 \geq \cdots$, Kohonen’s SOM algorithms iteratively define sets of cluster centers/prototypes $\left\{\mathbf{z}_{lm}^{j}\right\}$ for $j = 1, 2, \ldots$.

Kohonen's "online" SOM version

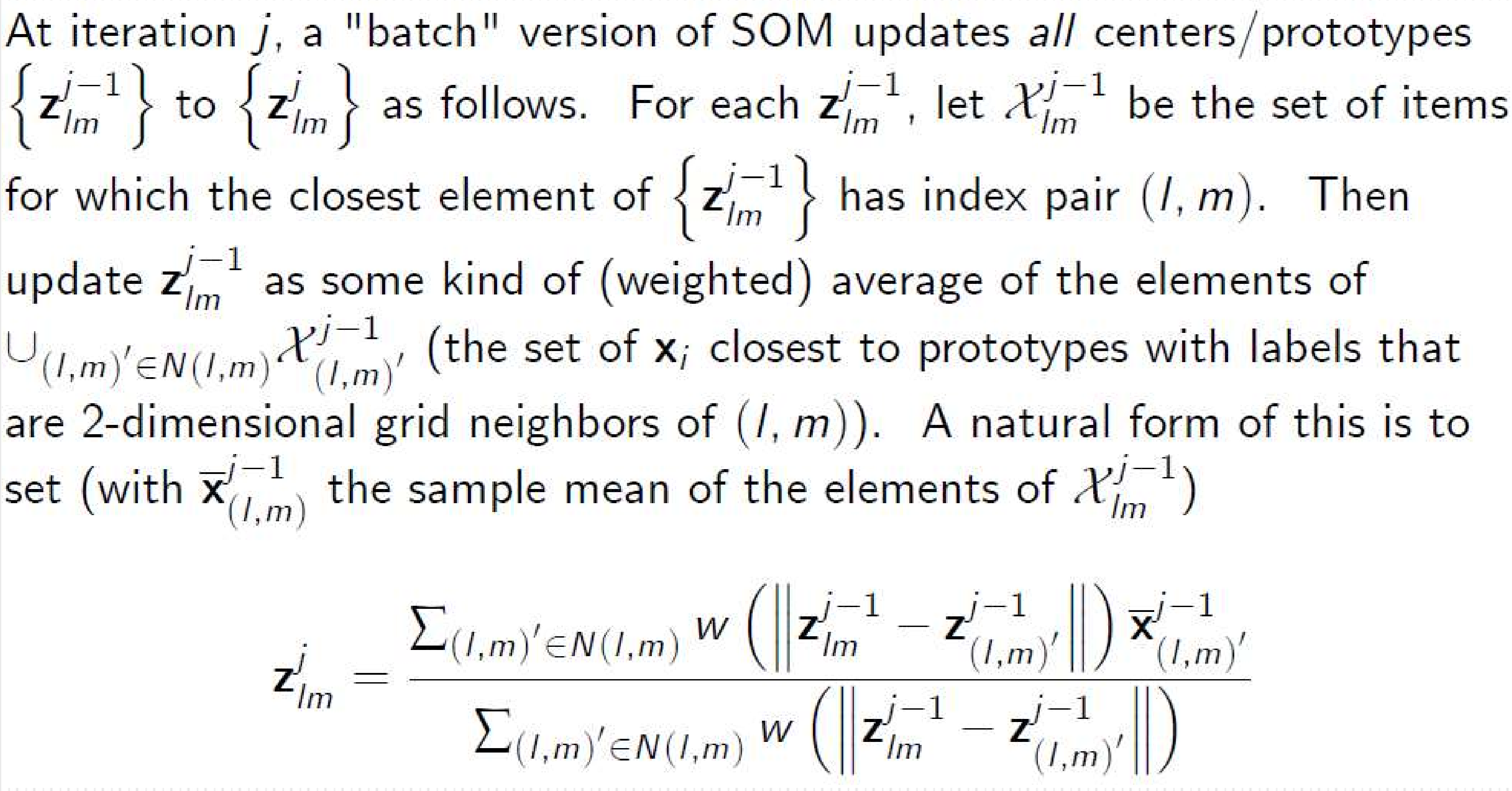


At iteration j, an "online" version of SOM selects (randomly or in turn from a randomly set ordering of the items) an item $\mathbf{x}^j$ and

1. identifies the center/prototype $\mathbf{z}_{lm}^{j-1}$ closest to $\mathbf{x}^j$ in $\Re^p$, call it $\mathbf{b}^j$ and its grid coordinates $(l, m)^j$ ($\mathbf{b}^j$ is the "BMU"/best-matching-unit),
2. adjusts those $\mathbf{z}_{lm}^{j-1}$ with index vectors belonging $N\left((l, m)^j\right)$ (close to the BMU index vector on the 2-dimensional grid) toward $\mathbf{x}^j$ by
$$\mathbf{z}_{lm}^{j} = \mathbf{z}_{lm}^{j-1} + \alpha_j w\left(\left\|\mathbf{z}_{lm}^{j-1} - \mathbf{b}^j\right\|\right)\left(\mathbf{x}^j - \mathbf{z}_{lm}^{j-1}\right)$$
(adjusting those centers different from the BMU potentially less dramatically than the BMU), and
3. for those $\mathbf{z}_{lm}^{j-1}$ with index pairs (l, m) not belonging $N\left((l, m)^j\right)$ sets
$$\mathbf{z}_{lm}^{j} = \mathbf{z}_{lm}^{j-1}$$

iterating to convergence.

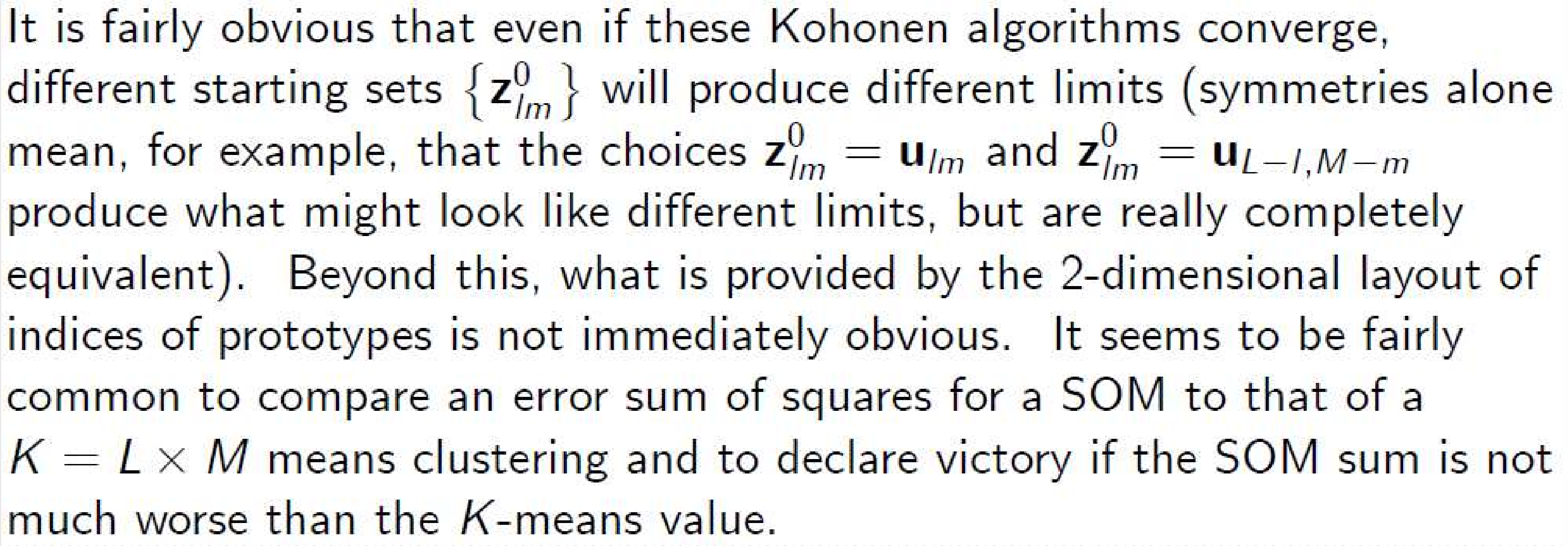
A “batch” SOM algorithm



At iteration j, a "batch" version of SOM updates *all* centers/prototypes $\left\{\mathbf{z}_{lm}^{j-1}\right\}$ to $\left\{\mathbf{z}_{lm}^{j}\right\}$ as follows. For each $\mathbf{z}_{lm}^{j-1}$, let $\mathcal{X}_{lm}^{j-1}$ be the set of items for which the closest element of $\left\{\mathbf{z}_{lm}^{j-1}\right\}$ has index pair (l,m). Then update $\mathbf{z}_{lm}^{j-1}$ as some kind of (weighted) average of the elements of $\cup_{(l,m)'\in N(l,m)}\mathcal{X}_{(l,m)'}^{j-1}$, (the set of $\mathbf{x}_i$ closest to prototypes with labels that are 2-dimensional grid neighbors of (l,m)). A natural form of this is to set (with $\overline{\mathbf{x}}_{(l,m)}^{j-1}$ the sample mean of the elements of $\mathcal{X}_{lm}^{j-1}$)

$$\mathbf{z}_{lm}^{j}=\frac{\sum_{(l,m)'\in N(l,m)} w\left(\left\|\mathbf{z}_{lm}^{j-1}-\mathbf{z}_{(l,m)'}^{j-1}\right\|\right)\overline{\mathbf{x}}_{(l,m)'}^{j-1}}{\sum_{(l,m)'\in N(l,m)} w\left(\left\|\mathbf{z}_{lm}^{j-1}-\mathbf{z}_{(l,m)'}^{j-1}\right\|\right)}$$

Comments on classical SOMs



It is fairly obvious that even if these Kohonen algorithms converge, different starting sets $\left\{\mathbf{z}_{lm}^{0}\right\}$ will produce different limits (symmetries alone mean, for example, that the choices $\mathbf{z}_{lm}^{0}=\mathbf{u}_{lm}$ and $\mathbf{z}_{lm}^{0}=\mathbf{u}_{L-l,M-m}$ produce what might look like different limits, but are really completely equivalent). Beyond this, what is provided by the 2-dimensional layout of indices of prototypes is not immediately obvious. It seems to be fairly common to compare an error sum of squares for a SOM to that of a $K=L\times M$ means clustering and to declare victory if the SOM sum is not much worse than the K-means value.

A more principled approach

Dissertation work of Rick Zhou takes a principled Bayesian modeling and decision-theoretic approach to the SOM objective. The following is an overview of his methodology.

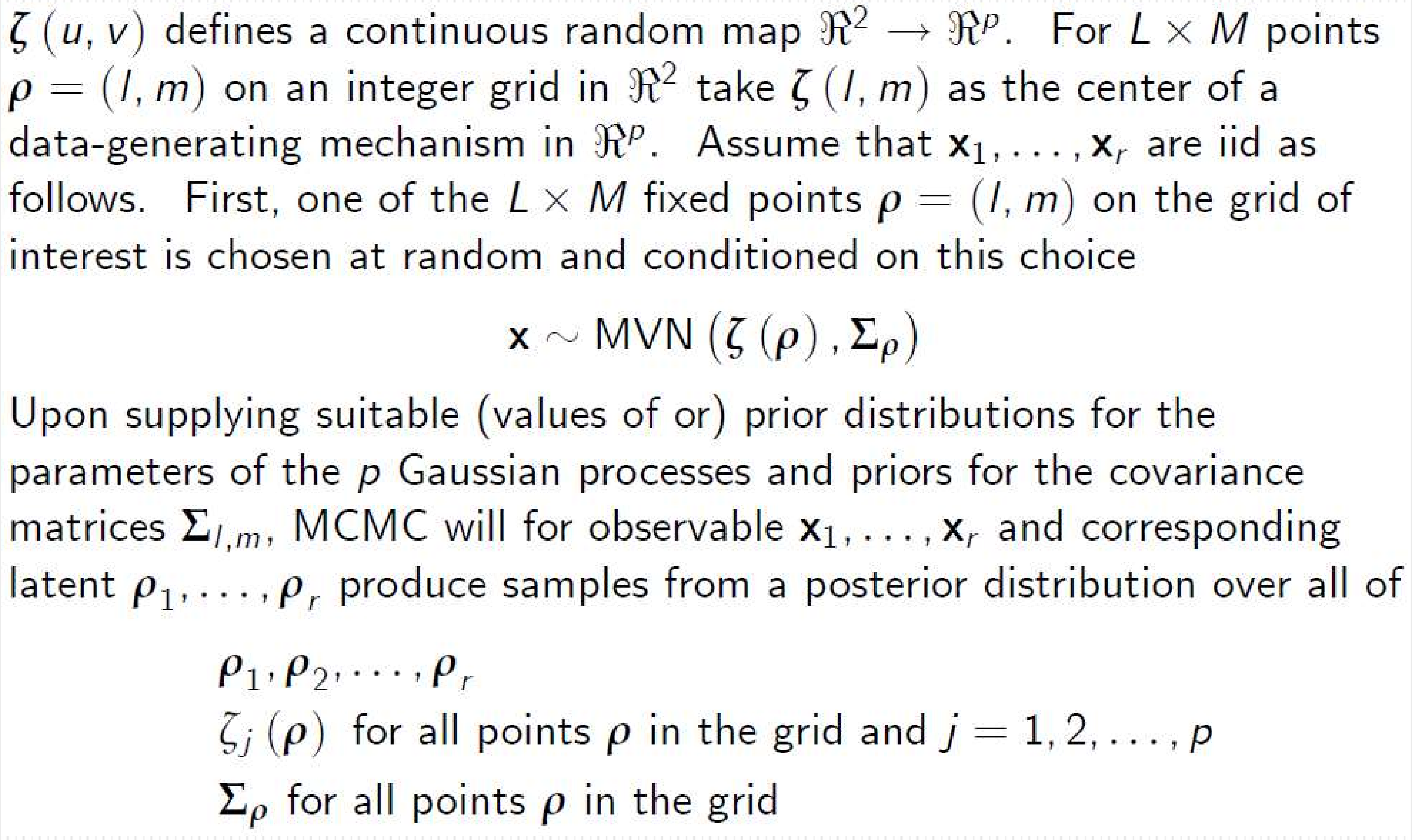
To develop a useful ("generative") model for $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r$ belonging to $\Re^p$, begin by defining p (one for each dimension of the data vectors) 0 mean Gaussian spatial processes

$$\zeta_1(u,v), \zeta_2(u,v), \ldots, \zeta_p(u,v)$$

and set

$$\boldsymbol{\zeta}(u,v) = \begin{pmatrix} \zeta_1(u,v) \\ \vdots \\ \zeta_p(u,v) \end{pmatrix}$$

More modeling



$\zeta(u,v)$ defines a continuous random map $\Re^2 \to \Re^p$. For $L \times M$ points $\boldsymbol{\rho} = (l,m)$ on an integer grid in $\Re^2$ take $\boldsymbol{\zeta}(l,m)$ as the center of a data-generating mechanism in $\Re^p$. Assume that $\mathbf{x}_1, \ldots, \mathbf{x}_r$ are iid as follows. First, one of the $L \times M$ fixed points $\boldsymbol{\rho} = (l,m)$ on the grid of interest is chosen at random and conditioned on this choice

$$\mathbf{x} \sim \mathrm{MVN}\left(\boldsymbol{\zeta}(\boldsymbol{\rho}), \boldsymbol{\Sigma}_{\rho}\right)$$

Upon supplying suitable (values of or) prior distributions for the parameters of the p Gaussian processes and priors for the covariance matrices $\boldsymbol{\Sigma}_{l,m}$, MCMC will for observable $\mathbf{x}_1, \ldots, \mathbf{x}_r$ and corresponding latent $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_r$ produce samples from a posterior distribution over all of

$\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \ldots, \boldsymbol{\rho}_r$

$\zeta_j(\boldsymbol{\rho})$ for all points $\boldsymbol{\rho}$ in the grid and $j = 1, 2, \ldots, p$

$\boldsymbol{\Sigma}_{\rho}$ for all points $\boldsymbol{\rho}$ in the grid

Computing and an objective

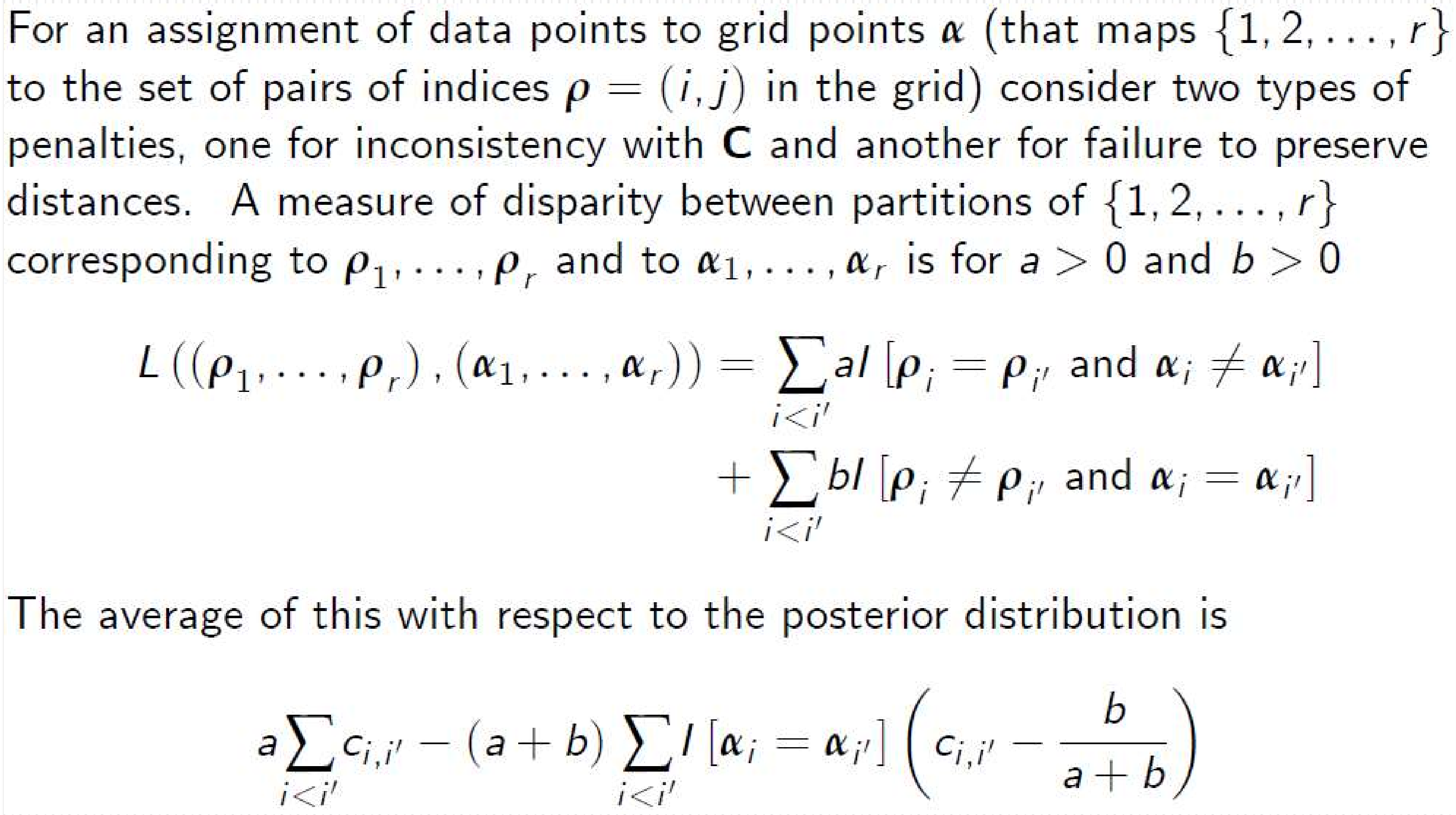
The grid points for the r cases, $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_r$, are of most interest. Two cases $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ belong to the the same cluster if $\boldsymbol{\rho}_i = \boldsymbol{\rho}_{i'}$. The MCMC provides relative frequencies that approximate posterior probabilities that case i and case i' belong together, $P\left[\boldsymbol{\rho}_i = \boldsymbol{\rho}_{i'}\right]$. That is, one obtains an estimate $\hat{\mathbf{C}}$ of the matrix

$$\underset{r\times r}{\mathbf{C}} = \left(P\left[\boldsymbol{\rho}_i = \boldsymbol{\rho}_{i'}\right]\right)_{\substack{i=1,2,\ldots,r\\ i'=1,2,\ldots,r}}$$

through MCMC relative frequencies and may seek an assignment of data points to grid points that

1. is consistent with $\mathbf{C}$, and
2. (at least locally) more or less preserves relative distances between clusters in $\Re^p$ in terms of distances between corresponding grid points in $\Re^2$.

Posterior average disagreement penalty



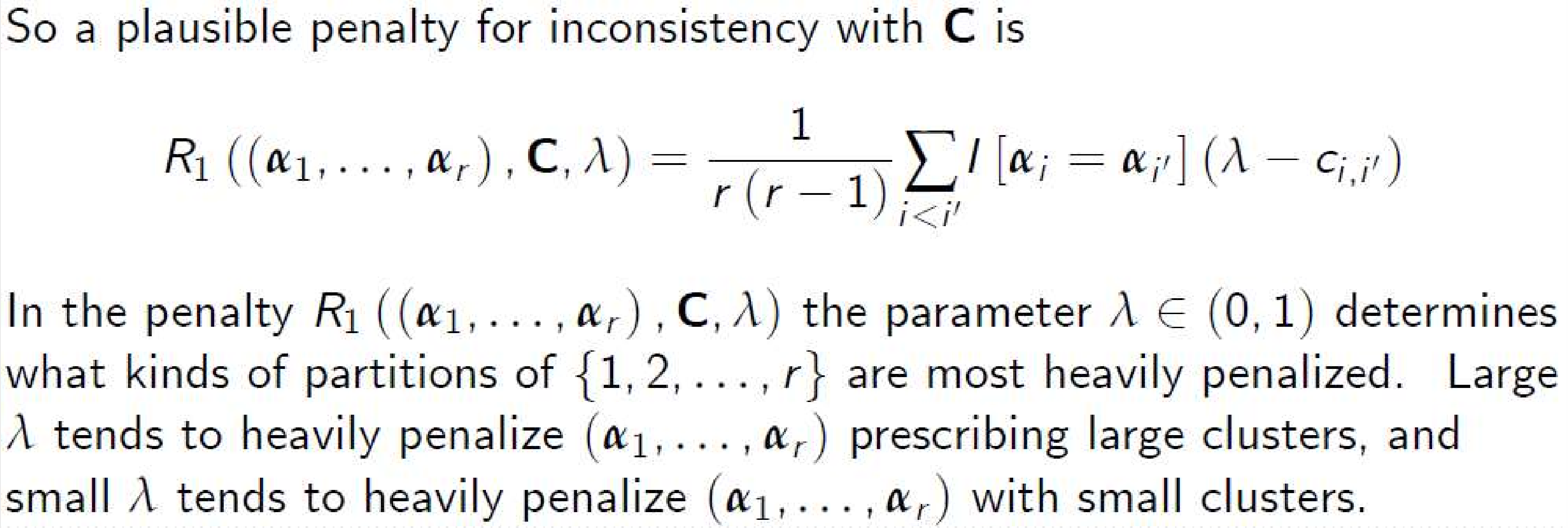
For an assignment of data points to grid points $\boldsymbol{\alpha}$ (that maps $\{1, 2, \ldots, r\}$ to the set of pairs of indices $\boldsymbol{\rho} = (i, j)$ in the grid) consider two types of penalties, one for inconsistency with $\mathbf{C}$ and another for failure to preserve distances. A measure of disparity between partitions of $\{1, 2, \ldots, r\}$ corresponding to $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_r$ and to $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r$ is for $a > 0$ and $b > 0$

$$\begin{aligned} L\left(\left(\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_r\right), \left(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r\right)\right) = & \sum_{i<i'} aI\left[\boldsymbol{\rho}_i = \boldsymbol{\rho}_{i'} \text{ and } \boldsymbol{\alpha}_i \neq \boldsymbol{\alpha}_{i'}\right] \\ & + \sum_{i<i'} bI\left[\boldsymbol{\rho}_i \neq \boldsymbol{\rho}_{i'} \text{ and } \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_{i'}\right] \end{aligned}$$

The average of this with respect to the posterior distribution is

$$a \sum_{i<i'} c_{i,i'} - (a+b) \sum_{i<i'} I\left[\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_{i'}\right] \left(c_{i,i'} - \frac{b}{a+b}\right)$$

Penalty for inconsistency with $\boldsymbol{C}$

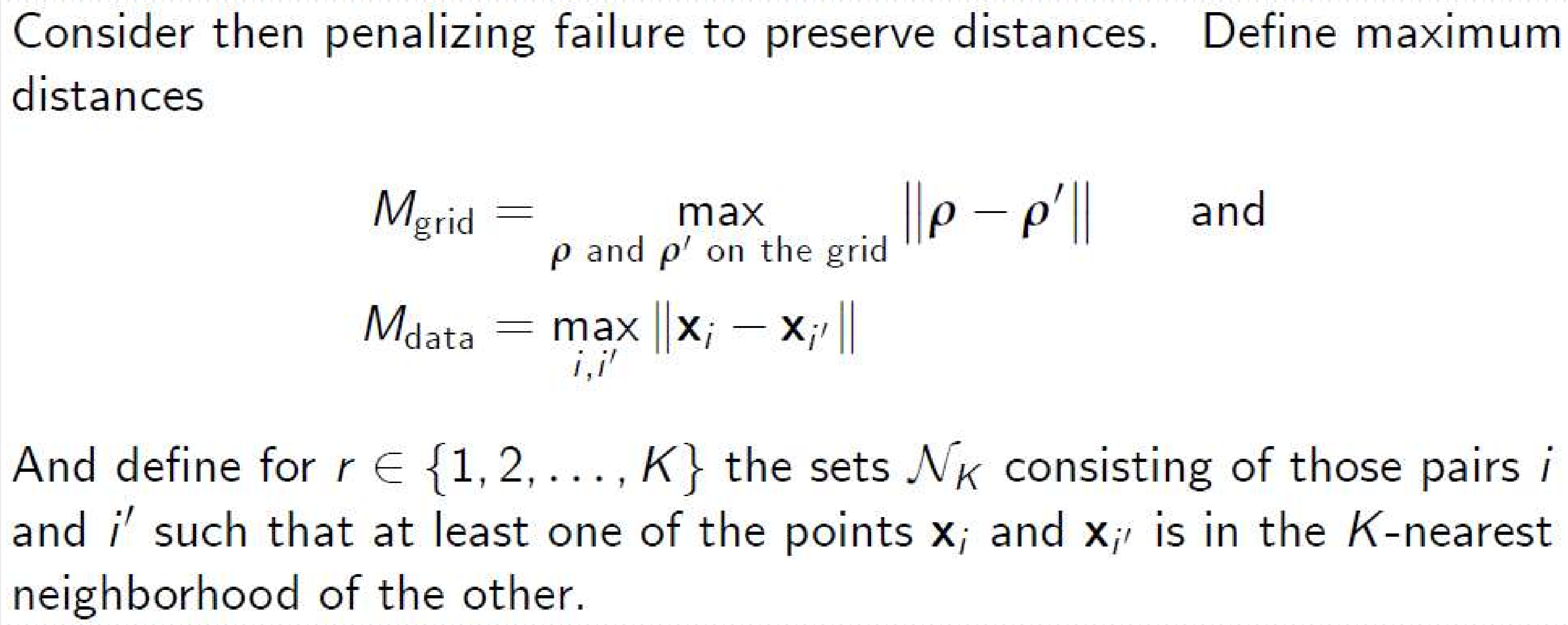


So a plausible penalty for inconsistency with $\mathbf{C}$ is

$$R_1\left(\left(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_r\right),\mathbf{C},\lambda\right)=\frac{1}{r\left(r-1\right)}\sum_{i<i'} I\left[\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_{i'}\right]\left(\lambda-c_{i,i'}\right)$$

In the penalty $R_1\left(\left(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_r\right),\mathbf{C},\lambda\right)$ the parameter $\lambda\in(0,1)$ determines what kinds of partitions of $\{1,2,\ldots,r\}$ are most heavily penalized. Large λ tends to heavily penalize $\left(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_r\right)$ prescribing large clusters, and small λ tends to heavily penalize $\left(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_r\right)$ with small clusters.

Penalty for failure to preserve distances



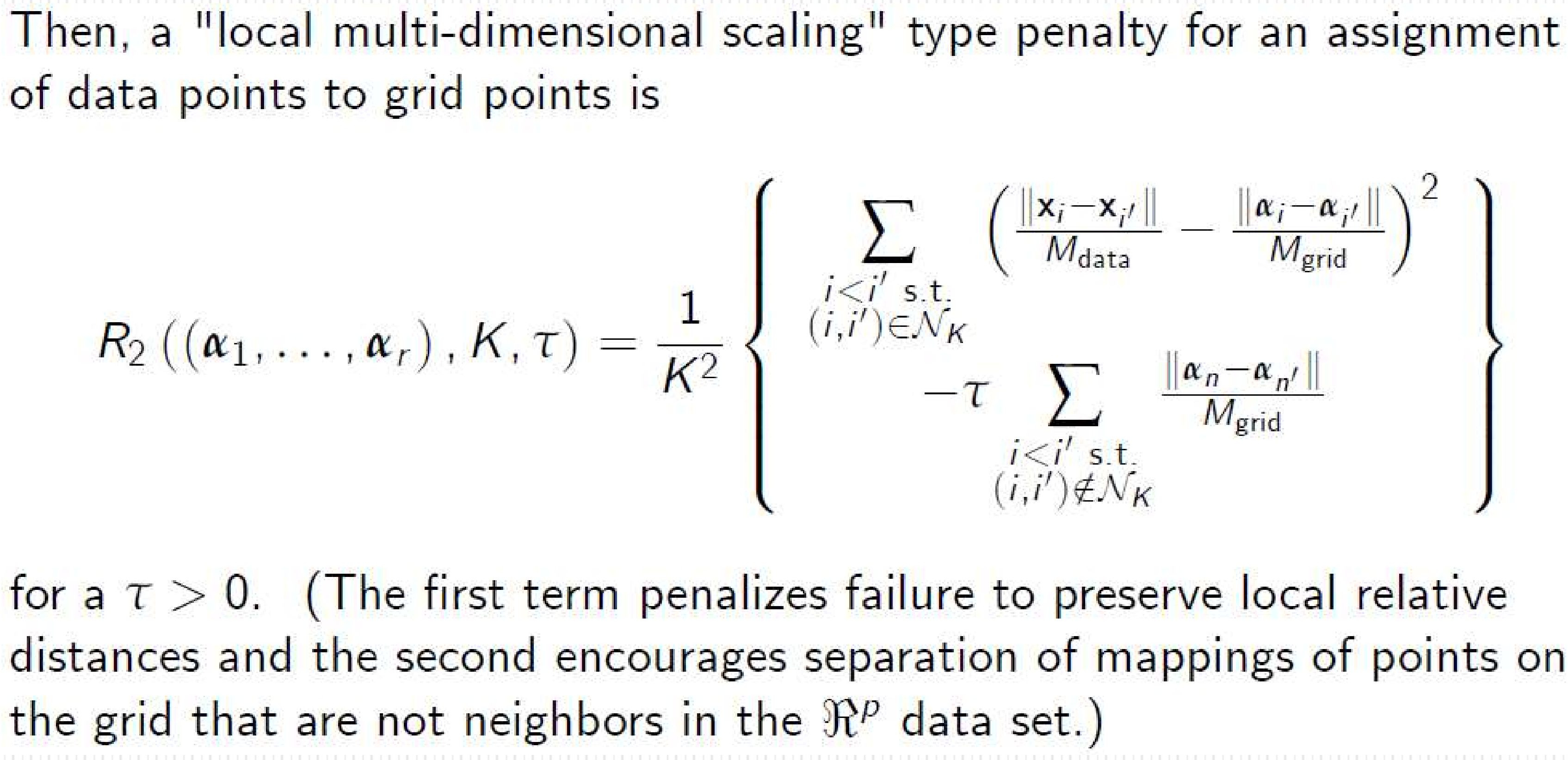
Consider then penalizing failure to preserve distances. Define maximum distances

$$M_{\text{grid}} = \max_{\rho \text{ and } \rho' \text{ on the grid}} \|\rho - \rho'\| \qquad \text{and}$$

$$M_{\text{data}} = \max_{i,i'} \|\mathbf{x}_i - \mathbf{x}_{i'}\|$$

And define for $r \in \{1, 2, \ldots, K\}$ the sets $\mathcal{N}_K$ consisting of those pairs i and i' such that at least one of the points $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ is in the K-nearest neighborhood of the other.

Penalty for failure to preserve distances cont.



Then, a "local multi-dimensional scaling" type penalty for an assignment of data points to grid points is

$$R_2\left(\left(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_r\right),K,\tau\right)=\frac{1}{K^2}\left\{\begin{array}{c}\displaystyle\sum_{\substack{i<i' \text{ s.t.}\\(i,i')\in\mathcal{N}_K}}\left(\frac{\|\mathbf{x}_i-\mathbf{x}_{i'}\|}{M_{\text{data}}}-\frac{\|\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_{i'}\|}{M_{\text{grid}}}\right)^2\\ \displaystyle-\tau\sum_{\substack{i<i' \text{ s.t.}\\(i,i')\notin\mathcal{N}_K}}\frac{\|\boldsymbol{\alpha}_n-\boldsymbol{\alpha}_{n'}\|}{M_{\text{grid}}}\end{array}\right\}$$

for a $\tau>0$. (The first term penalizes failure to preserve local relative distances and the second encourages separation of mappings of points on the grid that are not neighbors in the $\Re^p$ data set.)

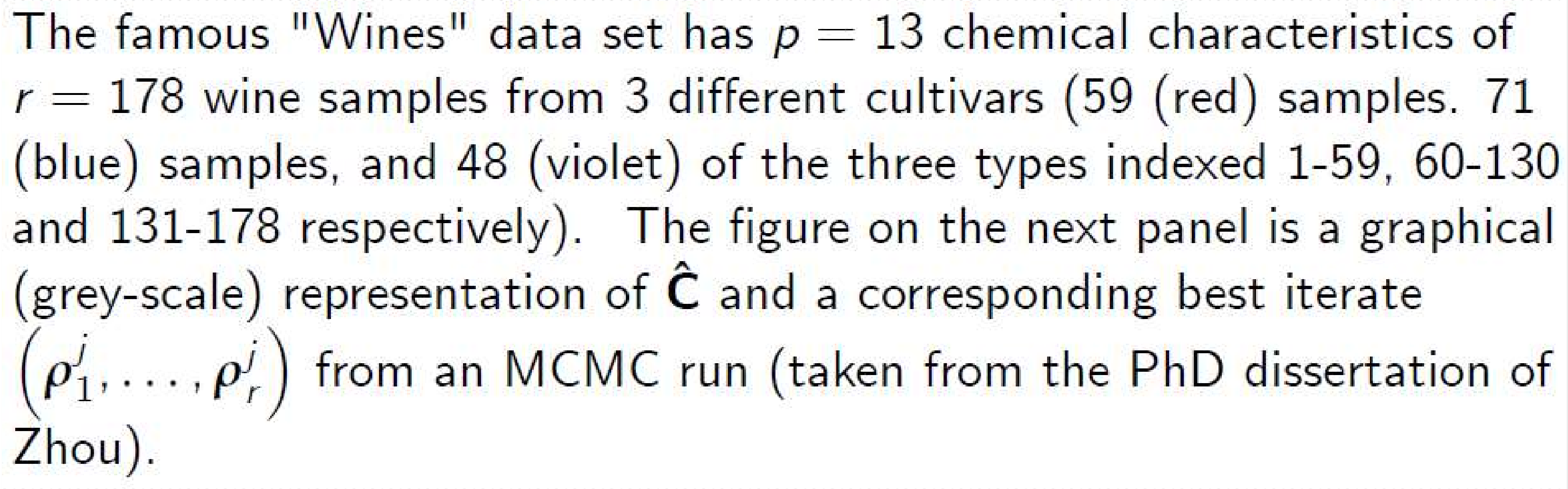
Approximately minimum posterior risk

So, a sensible risk/figure of merit for a map $\boldsymbol{\alpha}$ is for $\lambda > 0$

$$\begin{aligned} & R\left((\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r), \hat{\mathbf{C}}, \lambda, K, \gamma, \tau\right) \\ & = R_1\left((\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r), \hat{\mathbf{C}}, \lambda\right) + \gamma R_2\left((\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r), K, \tau\right) \end{aligned}$$

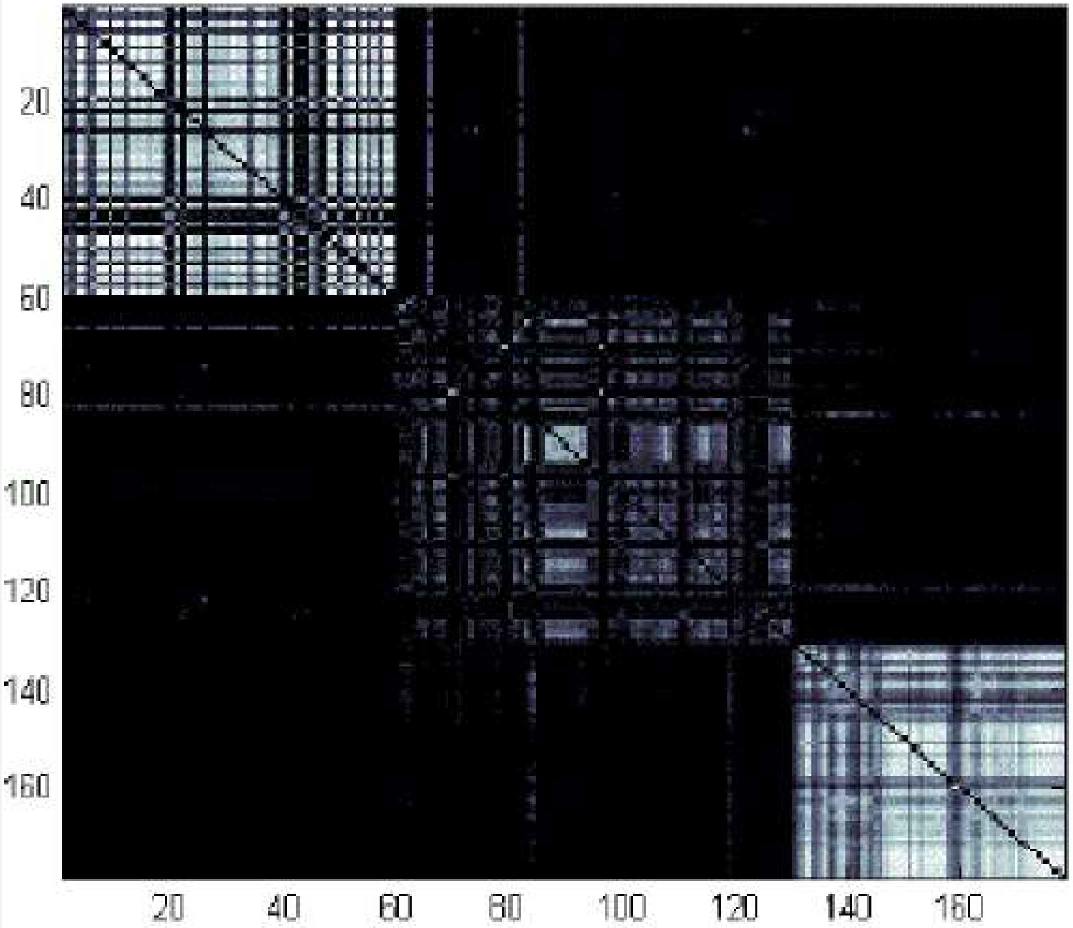
Exact optimization of $R\left((\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r), \hat{\mathbf{C}}, \lambda, K, \gamma, \tau\right)$ by choice of $(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r)$ is rarely computationally possible. What *is* possible and seems to work well is to make a long MCMC run (making the estimate $\hat{\mathbf{C}}$ reliable) and then look for an MCMC iterate $\left(\boldsymbol{\rho}_1^j, \ldots, \boldsymbol{\rho}_r^j\right)$ with the best value of $R\left(\left(\boldsymbol{\rho}_1^j, \ldots, \boldsymbol{\rho}_r^j\right), \hat{\mathbf{C}}, \lambda, K, \gamma, \tau\right)$. The Bayes model behind the MCMC tends to concentrate the posterior (and thus make iterates) in a manner consistent with the clustering and distance preservation goals of SOM.

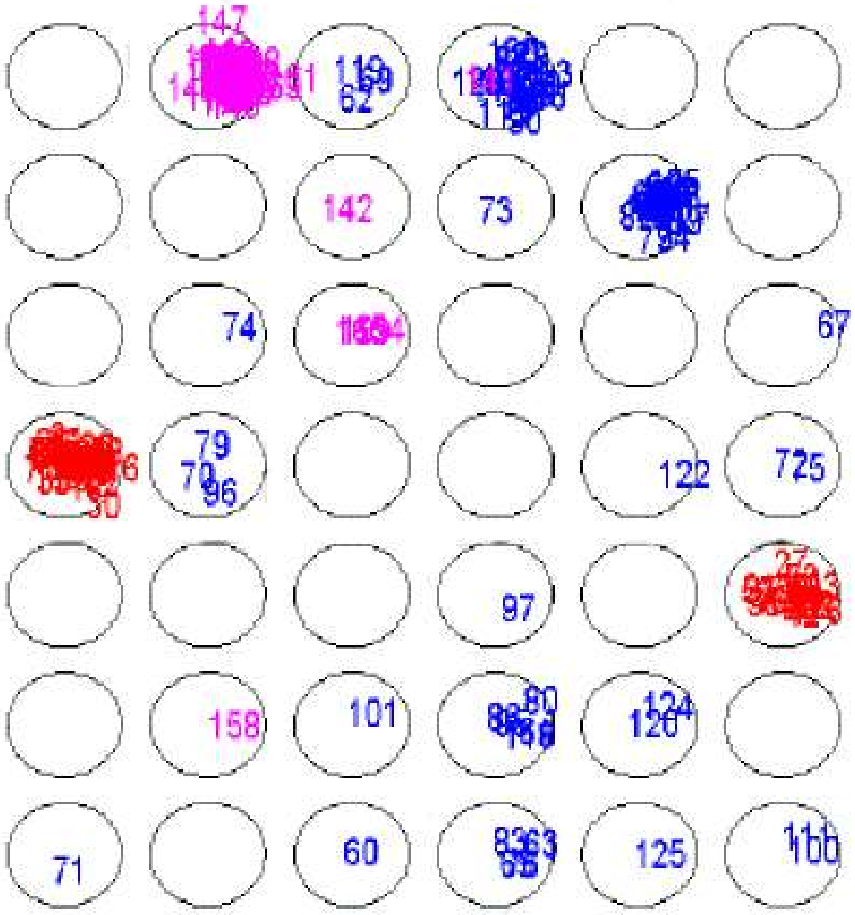
A real example



The famous "Wines" data set has $p = 13$ chemical characteristics of $r = 178$ wine samples from 3 different cultivars (59 (red) samples. 71 (blue) samples, and 48 (violet) of the three types indexed 1-59, 60-130 and 131-178 respectively). The figure on the next panel is a graphical (grey-scale) representation of $\hat{\mathbf{C}}$ and a corresponding best iterate $\left(\boldsymbol{\rho}_1^j, \ldots, \boldsymbol{\rho}_r^j\right)$ from an MCMC run (taken from the PhD dissertation of Zhou).

Bayes SOM for 178 wines


20
40
60
80
100
120
140
160
20
40
60
80
100
120
140
160


147
119
62
142
73
74
67
79
70
96
122
725
27
30
97
158
101
80
124
126
71
60
8363
125
111
100