Dimension Reduction (Mostly)

Bertrand Clarke[1]

[1] Department of Medicine
University of Miami

NDU 2011

Outline

The Need for Dimension Reduction

- Many data sets are a collection of n vectors $x = (x_1, ..., x_p)$ that we try to model as IID outcomes of $X = (X_1, ..., X_p)$.
- Standard theory imagined small vaues of p and medium to large values of n, with $p < n$.
- Now, frequently, p may be so large that no realistic sample size will ever be obtained for conventional inference.
- X may be a waveform, a huge graph, an image, or a document. Often data sets are multitype, meaning they combine qualitatively different classes of data.
- Thus, the complexity of the data (and model) means standard inference won't work. Thus, dimension reduction/variable selection becomes essential.

A Third Alternative

- Kernel methods have become popular because they don't do variable selection or feature extraction. They do data point selection!
- The model (such as it is) is found by using a kernel evaluated at a well-selected subset of the data points.
- The basic solution is

$$\hat{F}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) \quad \text{in a RKHS with kernel } K$$

- In effect, a kernel function $K(x, x')$ is a continuously parametrized collection of functions. For each fixed x', $K_{x'}(x)$ is a sort of basis element.
- If p is large, searching $\mathcal{F} = \{K_{x'}(x) | x \in I\!R^p\}$ to return a linear combination of a few elements can be useful.

Sparsity ctd.

- In a Representer Theorem solution, one can remove explanatory variables from x that do not contribute enough. This is a sort of double sparsity.
- It is standard to extract information from the covariates by themselves before bringing in the response.
- The goal here is to condense the information in the X_is into functions that have the information most relevant to modeling, regardless of the response.
- Then the information in the Ys can be used for variable selection on the extracted features.
- Variance-bias trade-off: Variable selection throws out variables that have too much variance for the amount of bias they eliminate. Feature extraction finds functions to get an even lower MSE from using fewer features.

Definition

- PCs, is one of a collection of techniques that extend linear regression by trying to identify underlying factors that explain a response.
- X does not to be normal, but PC's is a second moment technique.
- The idea behind PCs is to find a rotation of the original coordinate system in which to express the p-variate X_is so that each coordinate expresses as much of the variability in the X as a linear combination of the p entries can.
- Let $U = (U_1, ..., U_p)$ be a random vector and write $U = AX$ with $\text{Var}(X) = \Sigma$.
- Task: Find $A = (a_1, .., a_p)^T$ with $a_j = (a_{j,1}, ..., a_{j,p})$ such that

$$\forall j = 1, ..., p \;\; U_j = a_j^T X = \sum^{p} a_{j,k} X_k.$$

Properties

- We make $\text{Var}(U_j) = a_j^T \Sigma a_j$ is as high as possible subject to being uncorrelated with the other $U_j = a_j^T X$s; i.e., $\text{Cov}(U_j, U_k) = a_j^T \Sigma a_k = 0$ for $j \neq k$.
- The $\text{Var}(U_j)$s are decreasing.
- Thus, PCs are eigenvectors of the covariance matrix, ranked in order of the size of their eigenvalues.
- Linear combinations with high variance are the ones that affect the response the most.
- The variables least worth including are those with the smallest variability.
- If most of the variation comes from the first few PCs then it is enough to use them because the other linear combinations vary little from subject to subject.

PC Theorem Informal

- PCs can be peeled off one at a time from Σ by a sequence of optimizations. Start by finding $U_1 = a_1^T X$, where

$$a_1 = \arg \max_{\|a\|=1} \mathrm{Var}(a^T X). \tag{1}$$

- The a_1's from (1) is the direction in the X-space along which the variability is maximized; i.e., the eigenvector of Σ with maximal eigenvalue.
- To find U_2, or equivalently a_2, set

$$a_2 = \arg \max_{\|a\|=1, \mathrm{Cov}(a_1^T X, a^T X)=0} \mathrm{Var}(a^T X). \tag{2}$$

- Now, $U_2 = a_2^T X$. WLOG: $a_1 \perp a_2$ when $\lambda_1 > \lambda_2$. Later PCs are defined analogously; $a^T X$ is assumed uncorrelated with all the previous $a_j X$s.

PC Theorem, Formal

- Provided Σ is PD, it has a full set of p real eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p > 0$.
- The correct A has columns given by the eigenvectors e_1,..., e_p of Σ and the variances of the PCs are the eigenvalues.

Theorem: Let $\mathrm{Cov}(X) = \Sigma$ have eigenvectors e_1, .. ., e_p with corresponding eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p > 0$. Then:
(i) The j-th PC is $U_j = e_j^T X = e_{j,1} X_1 + \ldots + e_{j,p} X_p$ for $j = 1, ..., p$.
(ii) The variances of the U_j are $\mathrm{Var}(U_j) = e_j^T \Sigma e_j = \lambda_j$.
(iii) The covariances between the PCs are
$\mathrm{Cov}(U_j, U_k) = e_j^T \Sigma e_k = 0$.

- There are at least 2 proofs of this theorem. One is based on Lagrange multipliers the other is cleaner but rests on an auxiliary inequality.

Key Properties

- How do PC's characterize variability in X?
- Clause (ii) leads us to regard $\lambda_j / \sum \lambda_j$ as the proportion of variation of X explained by U_j.
- It is not just that PCs re-express the explanatory X so that the biggest contributions to variance can be identified.
- PCs permit sparsity in many cases because one can, for instance, regress on relatively few of the PCs or summarize data by using the PCs with, say, variances above a prechosen threshold.
- One can use thresholding arguments to get sparse representations for the PCs themselves.
- Two properties of PC's are helpful.

Theorem: Suppose $\mathrm{Cov}(X) = \Sigma$ and Σ has p eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$ with corresponding eigenvectors $e_j = (e_{j,1}, ..., e_{j,p})^T$.
(i) The sum of the variances of the U_js is

$$\sum_{j=1}^{p} \mathrm{Var}(U_j) = \lambda_1 + \cdots + \lambda_p = \sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{j=1}^{p} \sigma_{jj}.$$

(ii) The correlation between U_j and X_k is

$$\rho(U_j, X_k) = \frac{e_{jk}\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}.$$

- (ii) invites one to set some e_{jk}s to zero when they are small in absolute value, in addition to using only the first few PCs.
- The entries of X that get the highest weight in the first few normalized PC's might have more information.

Normal Case Interpretation

- If $X \sim N_p(\mu, \Sigma)$, then the density is constant on ellipses
$$(x-\mu)^T \Sigma^{-1} (x-\mu) = C^2.$$
Ellipses of this form have axes along $\pm C\sqrt{\lambda_j} e_j$.
- Set $\mu = 0$, and use the spectral representation of Σ as $\Sigma = \sum_{j=1}^{p} \lambda_j e_j e_j^T$ to give
$$C^2 = x^T \Sigma^{-1} x = \sum_{j=1}^{p} \frac{1}{\lambda_j} (e_j^T x)^2,$$
$e_j^T x$ is the component of x in the direction of e_j, the PCs.
- Writing $u_j = e_j^T x$ gives
$$C^2 = \frac{1}{\lambda_1} u_1^2 + \cdots + \frac{1}{\lambda_p} u_p^2.$$

- The PCs $U_j = e_j X$ lie in the directions of the axes of a constant-density ellipse. Any geometric point on the jth axis of the ellipse has coordinates in the x-frame proportional to e_j and in the u-frame (of the PCs) has coordinates proportional to a_j.
- It is often better to use the correlation matrix ρ than to use Σ. Also, since neither ρ nor Σ are usually available, it is important to be able to obtain PCs from data. These two variations are amenable to the same procedure as before. Second, the point of PCs is to use only the first few, say K. There are several ways to choose K.

Correlation PCs and Empirical PCs

- Instead of decomposing Σ, consider applying the method for deriving PCs to the correlation matrix for X, say ρ. Write $Z_j = (X_j - \mu_j)/\sqrt{\sigma_{jj}}$ so that $Z = (Z_1, ..., Z_p) = V^{-1/2}(X - \mu)$, where $V = \text{diag}(\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{pp}})$. Now, $EZ = 0$ and $\text{Cov}(Z) = V^{-1/2}\Sigma V^{-1/2} = \rho$, the correlation matrix of X. This puts the Z_js are on the same scale. Otherwise, X_js with larger scales will dominate an analysis.
- The PCs found from Z are not, in general, numerically the same as those found from X. Nevertheless, their forms and properties are the same and follow by proofs that are only slight modifications from before.

Same Theorem for Correlation Matrices

Theorem: Let (λ_1, e_1),..., (λ_p, e_p) be the eigenvalue, eigenvector pairs from ρ, with $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$. Then,
(i) The PCs of ρ are $U_j = e_j^T Z = e_j^T V^{-1/2}(X - \mu)$.
(ii) Variances are preserved in the sense that $\sum_j \text{Var}(U_j) = \sum_j \text{Var}(Z_j) = p$.
(iii) Correlations between U_j and Z_k are expressed in terms of the eigenvalues and eigenvectors of ρ, $\rho(U_j, Z_k) = e_{j,k}\sqrt{\lambda_j}$. $\square$

- Similarly, empirical forms of the PCs can be given. These result from using the PC procedure on estimates of Σ or ρ.
- Consider $\hat{\Sigma} = (1/n)\sum_{i=1}^{n} x_i x_i^T$ or $\hat{\rho} = \hat{V}^{1/2}\hat{\Sigma}\hat{V}^{1/2}$, although any other estimate of Σ or $\hat{\rho}$ could be used as well.
- The empirical PCs for X are $U_j = \hat{e}_1 X$, where $\hat{e}_j$ is the eigenvector corresponding to the jth largest eigenvalue $\hat{\lambda}_j$ of $\hat{\Sigma}$ or $\hat{\rho}$. The other properties of the PCs remain the same in both cases.
- Estimates of λ_j's and e_j's are asymptotically normal, centered at the true values with variances

$$2\lambda_j^2 \quad \text{and} \quad \lambda_j \sum_{k=1,k\neq j}^{p} [\lambda_k/(\lambda_k - \lambda_j)^2] a_k a_k^T,$$

respectively, where a_k is a vector of zeros with one in the kth entry.

How Many PCs?

- The point of PCs is dimension reduction. So, let K be the number of PCs to be retained in a model, $1 \leq K \leq p$. If $K = p$, there is no reduction. If $K = 1$, then a one-dimensional model, perhaps involving all the components of X, is the result. In most cases, $K \geq 2$ to capture all the useful variability in X.
- The three common ways to choose K are as follows. First, one can fix a proportion α of the variation to be explained by the PCs and let K be the smallest number of PCs required to achieve that. Choose K just large enough that

$$\frac{\hat{\lambda}_1 + \ldots + \hat{\lambda}_K}{\sum_{j=1}^{p} \hat{\lambda}_j} \geq \alpha.$$

This is somewhat arbitrary.

- Second, one can produce a scree plot. This is a graph of $\hat{\lambda}_j$ as a function of j. If the points are connected, then one can look for the knee in the curve, the point at which adding another PC adds relatively little explanatory power.
- Third, is to invoke a physical interpretation. One can look at the $\hat{\lambda}_j$s and recall that they represent the width of an ellipse of constant density, at least in the normal case. Thus, if a $\hat{\lambda}_j$ is small, the ellipse is narrow in the jth dimension and so the jth dimension may be neglected for physical reasons if the most important contributing components in $\hat{e}_j$ are known to be unimportant at that scale.
- By the same logic, if $\hat{\lambda}_1 \approx \hat{\lambda}_2$, it would be unreasonable to drop $\hat{\lambda}_2$ without dropping $\hat{\lambda}_1$ unless there were reason to believe that the components in $\hat{e}_2$ that had the greatest contribution were not helpful enough.

Factor Analysis

- Heuristically, the idea of factor analysis (FA) is to partition X into K strings of components $(X_1, ..., X_{p_1})$,...,$(X_{p_{K-1}+1}, ..., X_p)$ with $p = p_K$ with the property that the correlations within each string are high and the correlations between components from different strings are low. Summarize each string by a single 'factor'.
- Thus, FA is a generalization of PCs in which, rather than seeking a full-rank linear transformation with second-moment properties, one allows non-full-rank linear transformations.
- Consider modeling X as

$$X - \mu = \Lambda f + T,$$

where $EX = \mu$.

- More explicitly, for the $j = 1, ..., p$ entries of X, (3) is

$$X_j = \sum_{k=1}^{K} \lambda_{j,k} f_k + T_j + \mu_j.$$

- Here, Λ is a fixed $p \times K$ matrix of “loadings".
- The loadings $\lambda_{j,k}$ indicate how much X_j is affected by f_k; if several X_js have high values of $\lambda_{j,k}$ for a given factor f_k, then one may surmise that those X_js are redundant.
- The random $K \times 1$ vector f represents the common factors that underlie x.
- The random $p \times 1$ vector T represents the specific factors that underlie the particular experiment performed. Both T and f are unobservable.
- The goal is to explain the outcomes of X using fewer variables, the K unobserved factors in f, with $K << p$.

- Standardize the random quantities f and T

$$Ef = 0, ET = 0,$$

and the second moments are

$$\text{Cov}(f, T) = 0, \text{Cov}(T_j, T_{j'}) = 0 \text{ for } j \neq j', \text{Cov}(f) = Id_{K\times K}.$$

- Usually set $\text{Var}(T) = \text{diag}(\psi_1, ..., \psi_p) = \psi$.
- The assumptions give second-moment properties of X:

$$\text{Cov}(X, f) = \Lambda, \quad \text{Cov}(X_j, X_\ell) = \lambda_{j,1}\lambda_{\ell,1} + \ldots + \lambda_{j,K}\lambda_{\ell,K}$$

and

$$\text{Cov}(X_j, F_\ell) = \lambda_{j,\ell} \quad \text{and} \quad \Sigma = \Lambda\Lambda^T + \psi.$$

- $\Sigma = \Lambda\Lambda^T + \psi$ means that p variables correspond to K variables, and the $p(p-1)/2$ entries in the variance matrix are reduced to $K(K-1)/2 + p$ entries. Want $K << p$.
- It follows that, for $j = 1, ...p$,

$$X_j = \sum_{k=1}^{K} \lambda_{j,k} f_k + T_j - \mu_j$$

leading to

$$\sigma_{j,j} = \sum_{k=1}^{K} \lambda_{j,k}^2 + \psi_j = h_j^2 + \psi_j. \quad (3)$$

The h_j^2 is called the communality; it represents the part of the variance of X_j that comes from the underlying factors.

- ψ_i is the specific variance, from T_j, summarizing deviations that the common factors f_k, can't express.

Reduction to PC's

- FA reduces to PCs when $T = 0$ and the last $p - K$ eigenvalues of $\Sigma = \mathrm{Var}(X)$ are zero. In this case, write $\Sigma = \Gamma D \Gamma^T$ with the last eigenvalues in D zero, $d_{K+1}, ..., d_p = 0$, so that the upper left $K \times K$ block of D is $D_1 = \mathrm{diag}(d_1, ..., d_K)$ and only the upper left $K \times K$ block Γ_1 of Γ matters.
- Since PCs come from writing $U = \Gamma^T(X - \mu)$, it follows that $X - \mu = \Gamma U = \Gamma_1 U_1 + \Gamma_2 U_2$, where U_1 and U_2 are the first K and last $p - K$ components of U and Γ_2 is the lower right block of Γ. Thus, U_2 is trivial: It has mean and variance 0. So, $X - \mu = \Gamma_1 U_1$, which is

$$X = \Gamma_1 D_1^{1/2} D_1^{-1/2} U_1 + \mu_1.$$

Setting $\Lambda = \Gamma_1 D_1^{1/2}$ and $f = D_2^{-1/2} U_1$ gives the reduction.

- There are three sources of ambiguity in FA models: the choices of Λ, K, and f.
- In fact, Λ is only determined up to an orthogonal transformation. That is, let $K \geq 2$ and let V be any $K \times K$ orthogonal matrix, $VV^T = V^T V = I_{K \times K}$.
- The FA model can be written as

$$X - \mu = (\Lambda V)(V^T f) + T = \Lambda^* f^* + T \tag{4}$$

since $E(f^*) = 0$ and $\text{Cov}(f^*) = V^T \text{Cov}(f) V = I_{K \times K}$. The model is not identifiable.

- We get $\Sigma = \Lambda\Lambda + \psi = \Lambda^*\Lambda^* + \psi$. So, the communalities given by $\Lambda\Lambda = \Lambda^*\Lambda^*$ are also unchanged by V.

Finding Λ and ψ

- Extra conditions must be imposed to get unique estimates of Λ and ψ. In some cases, Λ can be purposefully rotated by V to make the results interpretable.
- More generally, estimating Λ and ψ is essential because they permit estimation of the factor scores f_j in f.
- The estimation procedures rest on $\hat{\Sigma}$. Let $\bar{x}$ denote the sample mean from $x_1, ..., x_n$, and denote the sample covariance matrix by $\hat{\Sigma}$ and the sample correlation matrix by $\hat{R}$. Setting $\hat{\sigma}_{j,j} = s_{j,j}$, write

$$\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}^T + \hat{\psi} \text{ and } \hat{\sigma}_{j,j} = \sum_{k=1}^{K} \hat{\lambda}_{j,k}^2 + \hat{\psi}_{j,j}.$$

An analogous expression can be written for $\hat{R}$.

Principal Factors

- The problem is to identify estimators $\hat{\Lambda}$ and $\hat{\psi}$ given K.
- PF's are related to PCs. In fact, when $\psi = 0$ or $K = p$, the eigenvector decomposition in PCs gives the FA representation.
- Basic idea is to start with the spectral decomposition of Σ and write

$$\begin{aligned}\Sigma &= \sum_{j=1}^{p} \lambda_j e_j e_j^T \approx \sum_{j=1}^{K} \lambda_j e_j e_j^T \\ &= (\sqrt{\lambda_1} e_1, \ldots, \sqrt{\lambda_1} e_K) \times (\sqrt{\lambda_1} e_1^T, \ldots, \sqrt{\lambda_1} e_K)^T \\ &= \Lambda\Lambda.\end{aligned}$$

- Note that the jth column of Λ is $\sqrt{\lambda_1} e_j$.
- That is, the jth factor loading comes from the jth PC and is exact if $K = p$, in which case $\psi = 0$.
- Of course, this is usually done on $\hat{\Sigma} = (1/n) \sum_{i=1}^{n} x_i x_i^T$ using $\hat{e}_j$ and $\hat{\lambda}_j$ to give $\hat{\Lambda}$ for a given K.
- In this case, $\hat{\psi}$ is usually not zero and it is typical to set $\hat{\psi}_j = s_{jj} - \sum_{k=1}^{K} \hat{\lambda}_{j,k}^2$ so $\text{diag}(\hat{\Sigma}) = \text{diag}(\hat{\Lambda}\hat{\Lambda} + \hat{\psi})$.
- The communalities are then $\hat{h}_j^2 = \sum_{k=1}^{K} \hat{\lambda}_{jk}^2$.
- The same procedure can be applied to the correlation matrix ρ.

Finding K

- As with PCs, the degree of dimension reduction depends on how small K can be chosen. This method is really only for principal factors.
- When the FA model is found using PCs, the natural way to evaluate how well it fits is to look at how good the approximation of Σ is. It can be shown that

$$\|\hat{\Sigma} - \hat{\Lambda}\hat{\Lambda} - \hat{\psi}\| \leq \hat{\lambda}_{K+1}^2 + \ldots + \hat{\lambda}_p^2, \tag{5}$$

  in which the norm is the sum of squares of the entries of the matrix. As K increases, the bound tightens.
- However, the goal is small K, meaning that the contributions from a small number of factors f_j to the sample variance should be large enough that the other factors can be neglected.

- The contribution to $s_{jj} = s_j^2$ from the first factor f_1 is $\hat{\lambda}_{j1}^2$.
- So, the contribution of f_1 to the total sample variance $\text{trace}(\hat{\Sigma}) = s_{11} + \ldots + s_{pp}$ is

$$\sum_{j=1}^{p} \hat{\lambda}_{j1}^2 = (\sqrt{\hat{\lambda}_1}\hat{e}_1)^T(\sqrt{\hat{\lambda}_1}\hat{e}_1) = \hat{\lambda}_1,$$

  the $(1,1)$ entry of $\hat{\Lambda}\hat{\Lambda}^T$, where $\hat{\Lambda} = (\sqrt{\hat{\lambda}_1}\hat{e}_1), \ldots, (\sqrt{\hat{\lambda}_p}\hat{e}_p)$, and this holds for 2, 3,..., p.
- The proportion of the total sample variance attributable to the jth factor is $\hat{\lambda}_j / \sum_{j=1}^{p} s_{jj}$ and $\hat{\lambda}_j / p$ when factor analysis is applied to $\hat{\Sigma}$ or $\hat{\rho}$, respectively. Since the eigenvalues are decreasing, the way we chose K for PC's cantinue to apply.

Estimating Factor Scores

- The choice of K gives the degree of dimension reduction from p, but it remains to convert the p-dimensional data points x_i to new points $\hat{f}_i$, called factor scores, in K dimensions.
- Note that the choice of K, Λ, and ψ is determined by all the x_is, so adding another data point may change the model.
- The basic problem is that there are n known values, the x_is, but $2n$ unknown values, the ϵ_is and the f_is. One way to convert x_is to factor scores (estimates) is by weighted least squares. Start with fixed values for Λ, μ, and ψ and treat T as if it were an error term, ϵ. So, the model is $X_i - \mu = \Lambda f_i + \epsilon_i$ for $i = 1, \ldots, n$ and consider determining f_i for x_i.

- The least squares strategy for overcoming the indeterminacy is as follows. Observe that the sum of squares due to error is

$$SSE = \sum_{j=1}^{p} \epsilon_j^2/\psi_j = \epsilon\psi^{-1}\epsilon = (X-\mu-\Lambda f)^T\psi^{-1}(X-\mu-\Lambda f).$$

- Minimizing it gives $f_{min} = (\Lambda^T\psi^{-1}\Lambda)^{-1}\Lambda^T\psi^{-1}(x-\mu)$. So, for $i = 1, ..., n$, when Λ and ψ are estimated by the ML method (and $\hat{\Lambda}\hat{\psi}^{-1}\Lambda = \hat{\Delta}$), it is natural to set

$$\hat{f}_i = (\hat{\Lambda}^T\hat{\psi}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}^T\hat{\psi}^{-1}(x_i - x).$$

- When the principal factors are used, the $\hat{\psi}$ drops out: The results are

$$\hat{f}_i = (\hat{\Lambda}^T\hat{\Lambda})^{-1}\hat{\Lambda}^T(x_i - \bar{x})$$

for Σ, which can be recognized as the first K scaled PCs

Examples

- Consider the following data analyzed in Abdi (2003). There are five wines and seven measurements are made on each by a panel of experts: how pleasing it is, how well it goes with meat or dessert, its price, sweetness, alcohol content and acidity.

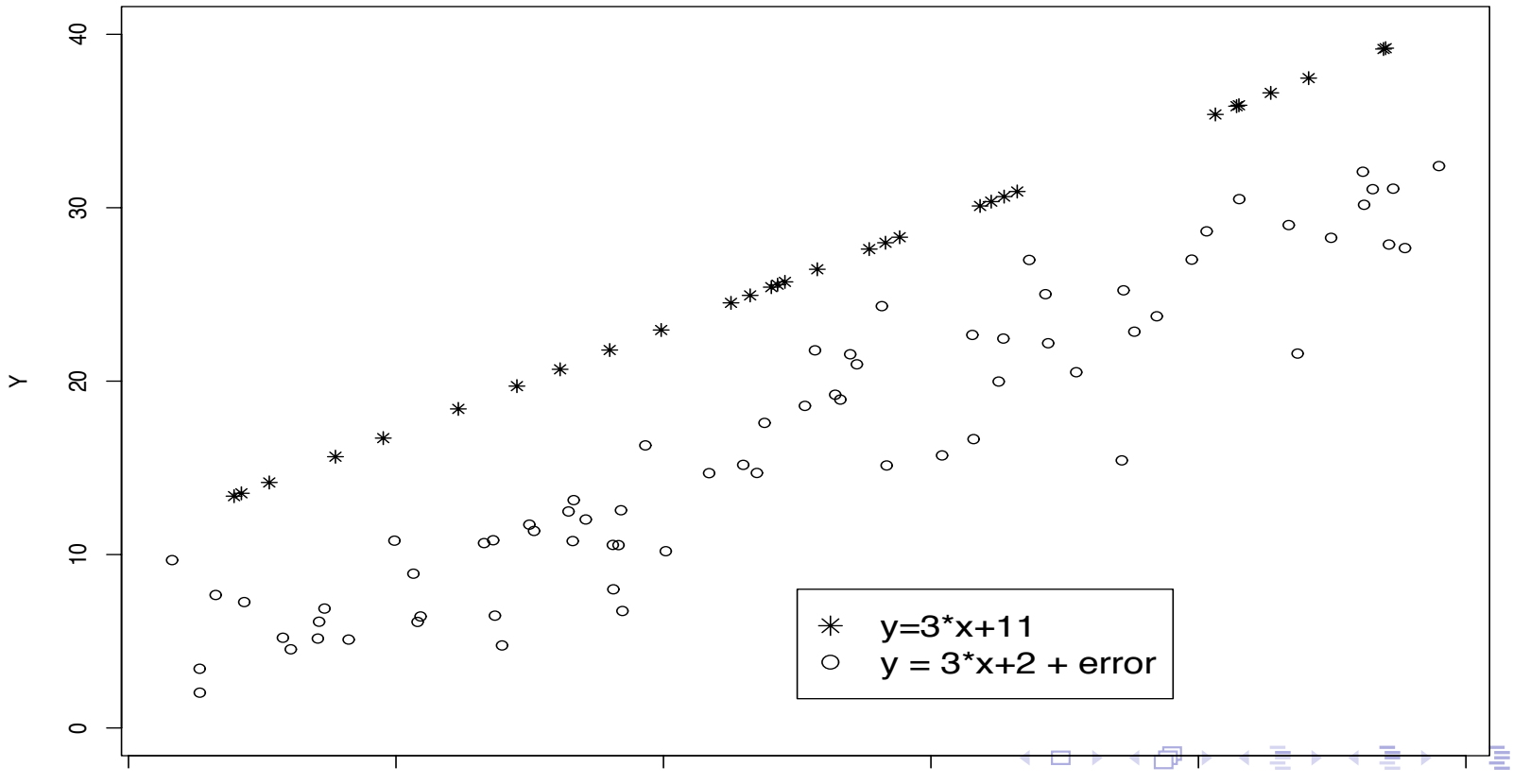
Wine	Hedonic	Meat	Dessert	Price	Sugar	Alcohol	Acid
1	14	7	8	7	7	13	7
2	10	7	6	4	3	14	7
3	8	5	5	10	5	12	5
4	2	4	7	16	7	11	3
5	6	2	4	13	3	10	3

- The principal component solution indicates that 93.90 % percent of the variance is explained by the first two components.
- The matrix of factor loadings can be found using factanal. The 2×7 table of factor loadings is given below:

	Hedonic	Meat	Dessert	Price	Sugar	Alcohol	Acid
1	-0.40	-0.45	-0.26	0.42	-0.05	-0.44	-0.45
2	0.117	-0.117	-0.597	-0.31	-.72	0.06	0.09

- When you have complex data and no idea what models might be appropriate take a look.
- There may in fact be a very intelligible model; however, teasing it out from the data in the absence of physical knowledge may be exceedingly difficult.
- Visualization, like dimension reduction and clustering, can be regarded as a collection of search strategies to find some regularity in the data strong enough that modeling can say something about it.
- Consider linear regression data (Y_i, X_i) for $n = 97$, with 27 points on a straight line and 70 points generated with noise. There are two subsets. However, naive regression will miss the subsets, defaulting to an average solution.

Two Clusters



What to do?

- It would be nice to be able to cherry-pick.
- Maybe use a clustering technique to find (hopefully two) clusters. Then, choose the largest cluster and fit a regression model.
- Provided the model fits well, remove outliers from the cluster according to some reasonable criterion and search the other clusters for points that fit the model well, putting them into the first cluster if they do not worsen fit much. Reclustering the remaining points and repeating the procedure might reveal the two-cluster structure.
- In general, the goodness-of-fit measure should not depend on the sample size or p (e.g. adjusted R^2); (ii) modeling within a cluster should trade-off sample size and number of variables, and (iii) K is uncertain.

- If this kind of procedure were used with the data and the first cluster with 70 points had been found, the graph of most reasonable assessments of fit would look something like the next figure as data points from the smaller cluster were added. This plot is based on using R^2 for fitness with the data from before.
- The total sample size used here is 80; the first 70 observations are from the noisy cluster and the last 10 observations were generated exactly on the line, as indicated by the knee in the curve.
- Although this procedure is not formal, it does accurately reveal the structure of the data. In principle, one could propose a model for how adding wrong model data affects statistics such as R^2, generate a curve with a knee and test whether the points added beyond the knee were sufficiently different to reflect a different model

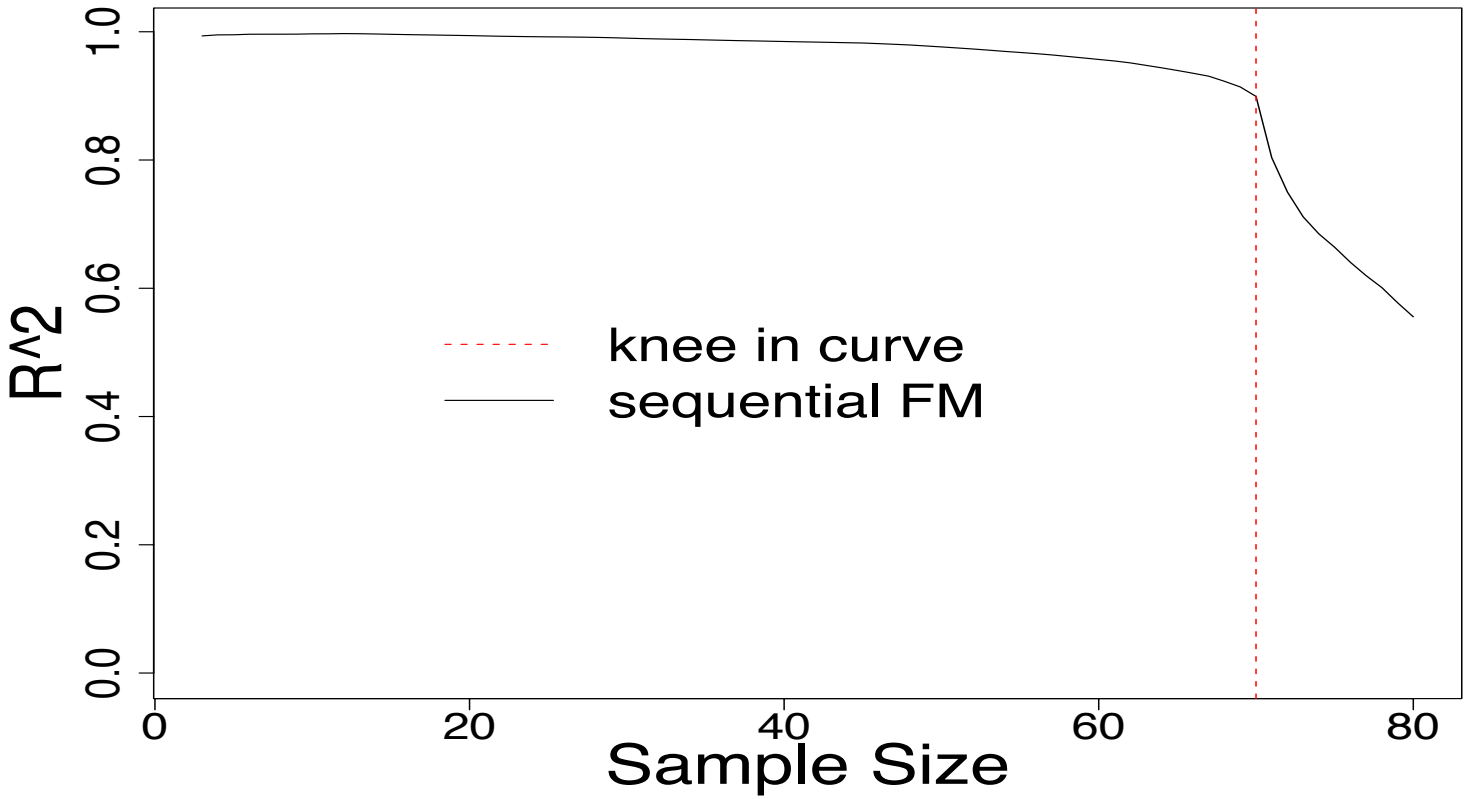


Figure: How accumulating data within a model affects fit. The first 70 data points fit. Past 70 the data points were from the smaller cluster.

Warning

- Making diagrams from data uses information. This means that the probability of type I error in subsequent testing will be larger than the stated α and the actual standard errors of estimators will be larger. Inferences after data snooping will necessarily be much weaker.
- Visualization, dimension reduction, clustering ise information faster than calculating individual statistics and really should be done selectively if downstream inference is of great importance.
- OK for presenting results or to search for structure to model. But downstream formal inferences are affected.

Elementary Visualization

- Here are several common ways to represent data with little or no processing. These techniques are most useful when the dimension is between 4 and, say, 20 or so.
- A profile in p dimensions is a representation of a vector of the form $(x_1, ..., x_p)$ in which the values x_j are plotted adjacent to each other. This can be presented as a bar graph with p bars on a common axis or as a polygonal line.
- A star in p dimensions is a representation of $(x_1, ..., x_p)$ in which the values x_j are plotted on axes drawn from a center point. Any two adjacent axes have the same angle between them. The values x_j are noted on the axes and then connected to form a p-gon.

- Doing this for a data set gives n p-gons that may reveal patterns, depending on the ordering of the x_js.
- Using data on 17 classes of household expenditures from nine Canadian cities for 2006, the next figure shows one star with 17 points for each city. Household expenditures means dollars per year spent on food, shelter, clothing, and so forth.
- The figure after it shows the profiles for the cities plotted on the same axis. The four peaks are suggestive, but a little misleading if read too closely: The four largest peaks at 2, 6, 9, and 15 correspond to shelter, transportation, recreation, and taxes. However, the first variable, food, should be a peak: It is higher than all the other variables, except for shelter, transportation, and taxes, which is the biggest expense.

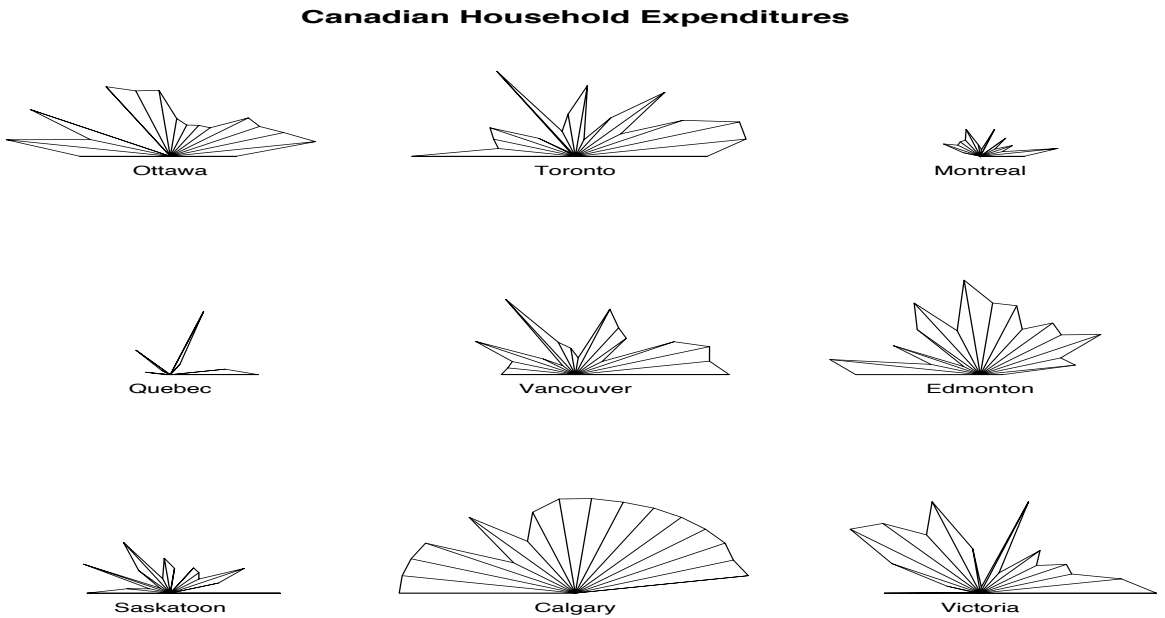


Figure: Calgary's star fans out the most uniformly over the classes, while Quebec City's star has only five points of any real size; they are food, shelter, transportation, insurance and pensions, and taxes.

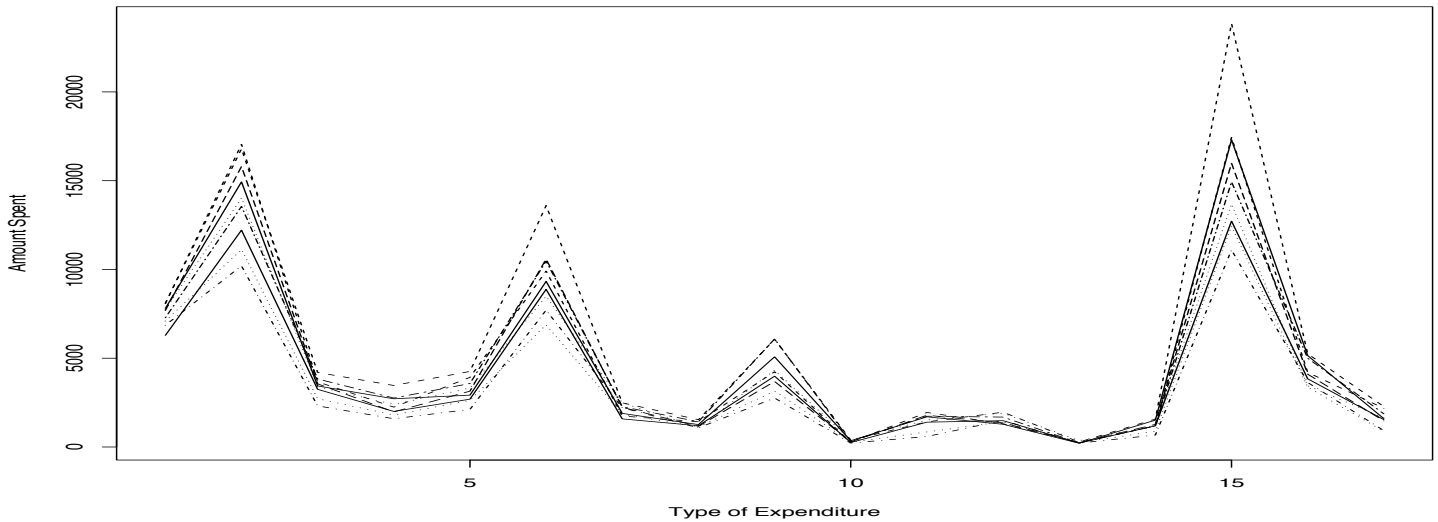


Figure: The lower panel shows the profiles. There are four peaks, but the point is that across Canadian cities the distribution of expenditures is fairly similar.

Heatmap

- A heatmap is a matrix of values that have been color coded, usually so that higher values are brighter and lower values are darker.
- Usually the rows and columns are grouped so that those in the same group are next to each other; this leads to figures comprised of homogeneous rectangles.
- Heatmaps are often good for presenting data once they have been analyzed, but often do not reveal much because the patterns tend to be weak.
- Heatmap for Motor Trend 1974 data. The variables were: mpg; number of cylinders; disp, displacement (cu.in.); hp; rear axle ratio; wt; qsec, 1/4 mile time; vs; transmission (0 = automatic, 1 = manual); number of forward gears; and number of carburetors.

- Both the models and the measurements have been clustered. Roughly, the models of cars are in three clusters: The bottom cluster (Duster to Maserati) consists of cars that are heavy or have big engines; the top cluster (Honda Civic to Toyota Corona) consists of lighter cars with smaller engines; and the middle cluster (Valiant to Mercedes 450SL) is in between.
- The clustering on the measurements on the cars is less clear: The pair at the bottom are measures of power and the next two are measures of performance, but it is unclear what the block of seven (cyl to gear) represents.
- The heatmap itself shows a clear dividing line between the Honda Civic and the Mercedes 450SL: In each row, if the left part is light, the right part is dark, or vice versa. Maybe the middle cluster has more in common with the right cluster than the left?

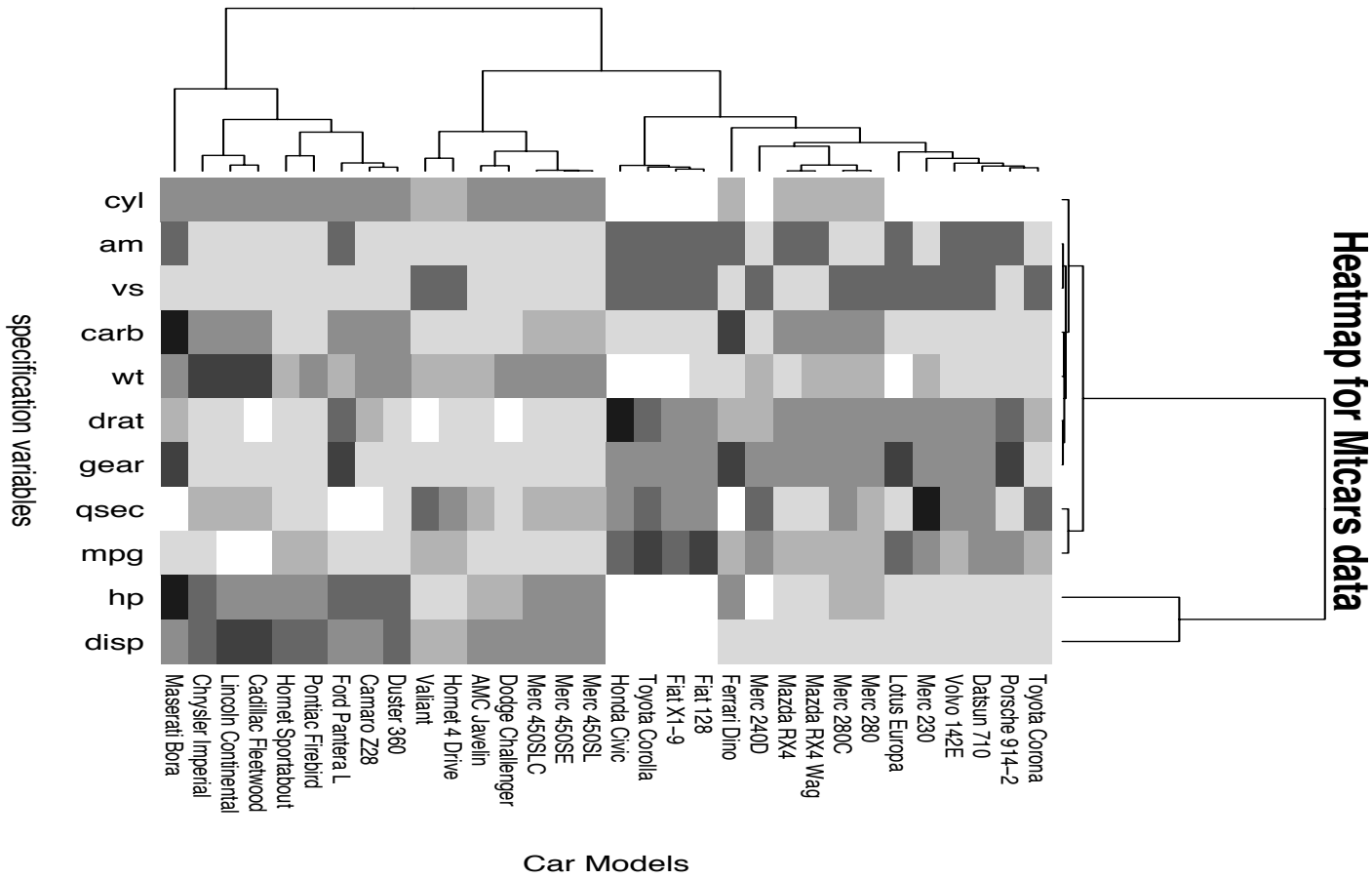
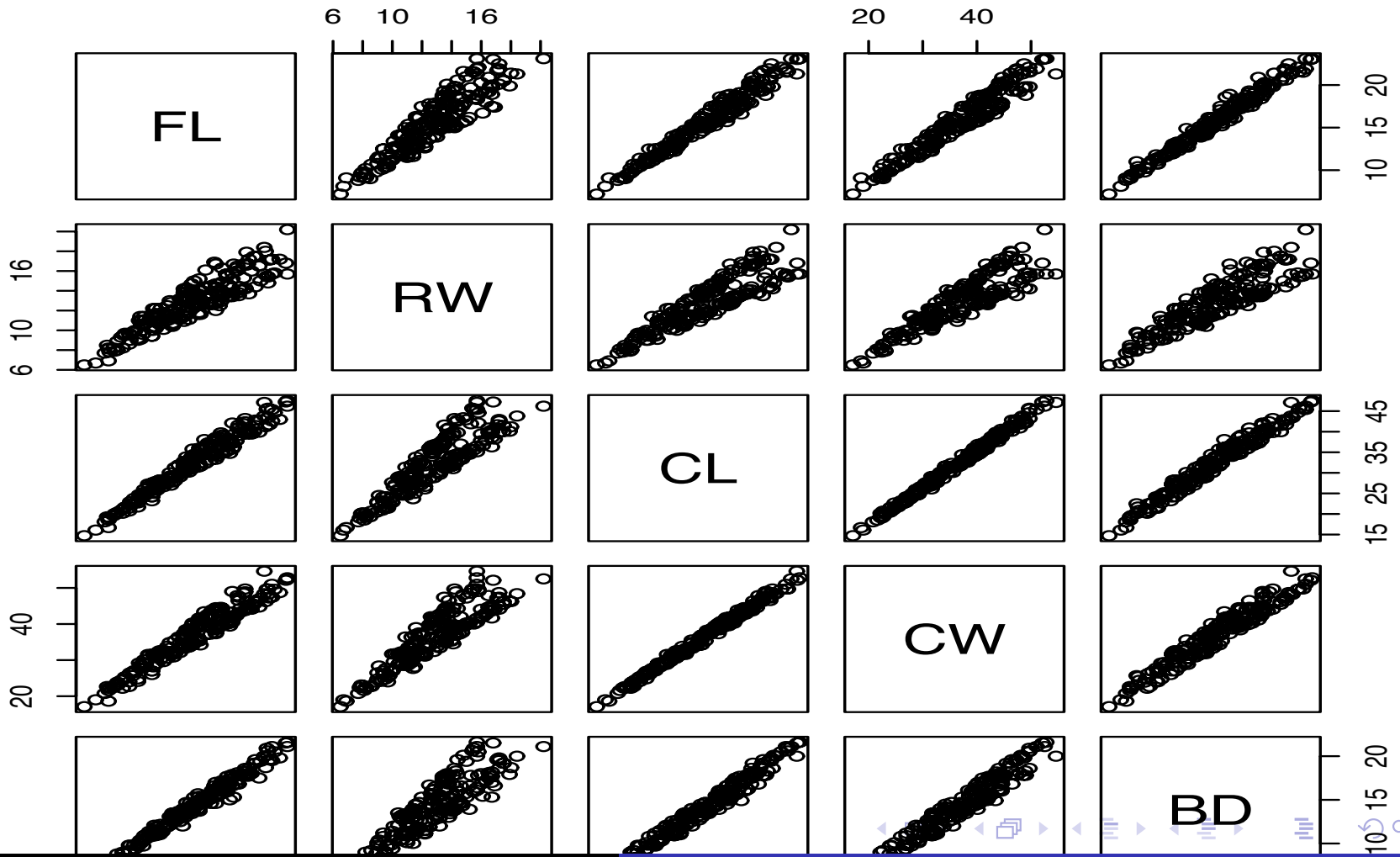


Figure: Heatmap with clustering on the models and variables done separately. Darker regions correspond to higher values.

Projections

- p dimensional data are mostly understood by looking at their projections onto two dimensions.
- So, we want to find the directions along which the projection of the data will reveal its features.
- Consider projecting p-dimensional points $(x_{i,1}, ..., x_{i,p})$ with a $p \times p$-dimensional idempotent matrix D.
- A value x_{i,j_0} may be an outlier in one plane but not in another. (Think of a curve in the horizontal plane and one point several units above it.)
- One way to search for outliers is through all pairwise scatterplots from projections into the coordinate planes: $p(p-1)/2 - p$ scatterplots of $(x_j, x_{j'})$ for $j' > j$, the upper triangle of a matrix of scatterplots.

- Another way to do this is to spin the data. The idea is to project the data points into a three-dimensional subspace and then rotate the projections.
- Rotation can be interactive (user controlled) or automated. Systematically doing this so that all representative projections of the data are seen is called a Grand Tour.
- Watching the rotations in continuous time reveals the context of informative projections as well as the projected points themselves.
- Consider the Australian crab data available from `www.ggobi.org/book/`. The data consist of 200 measurements on a sample of crabs from Australia. Each measurement is five-dimensional: frontal lobe length, FL; rear width, RW; carapace length, CL; carapace width, CW; and body depth, BD.


FL
RW
CL
CW
BD
6
10
16
20
40
15
25
35
45

- The data are five dimensional, one can load them into the ggobi or rggobi visualization system, which can be downloaded freely from `www.ggobi.org`.
- GGobi can generate Grand Tours. Doing this, one can watch the way the data forms change shape as the perspective is rotated. GGobi can be paused at interesting projections; it also gives the unit vectors defining each projection.
- First, the left [anel was found by watching the Tour and stopping it at a clearly delineated shape.
- Then the picture was rotated by dragging the cursor the right way. This generated the other two panels on the right.
- The next step would be to color one of the arms of the vee in the left and do the rotation again to see how the points changed their relative positions. Doing this permits separation of each arm of the vee

- The first right panel shows that the circles and triangles form the lower arm of the vee and the squares and diamonds form the upper arm.
- The rotation to the second panel on the left brings the circles alone to the top, the squares and triangles to the middle region, and the diamonds alone to the bottom.
- Continuing the rotation, the third panel on the left shows that the circles and squares are on the top and the triangles and diamonds are on the bottom.
- It is left as an exercise to use GGobi to find a direction (down the center) in which the cones collapse into four blobs, one for each species-sex pair.

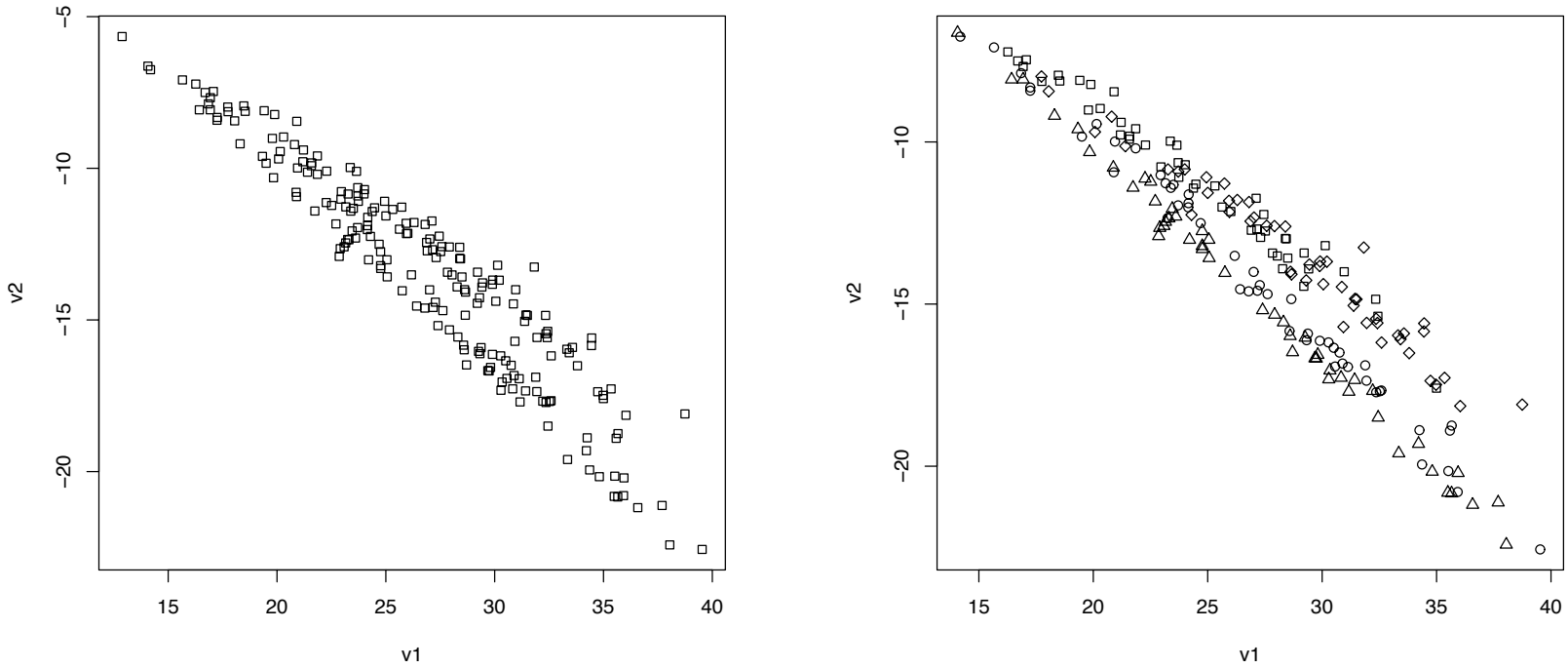


Figure: The left panel shows a view found from GGobi. The right panel is the same view but with the cones labeled.

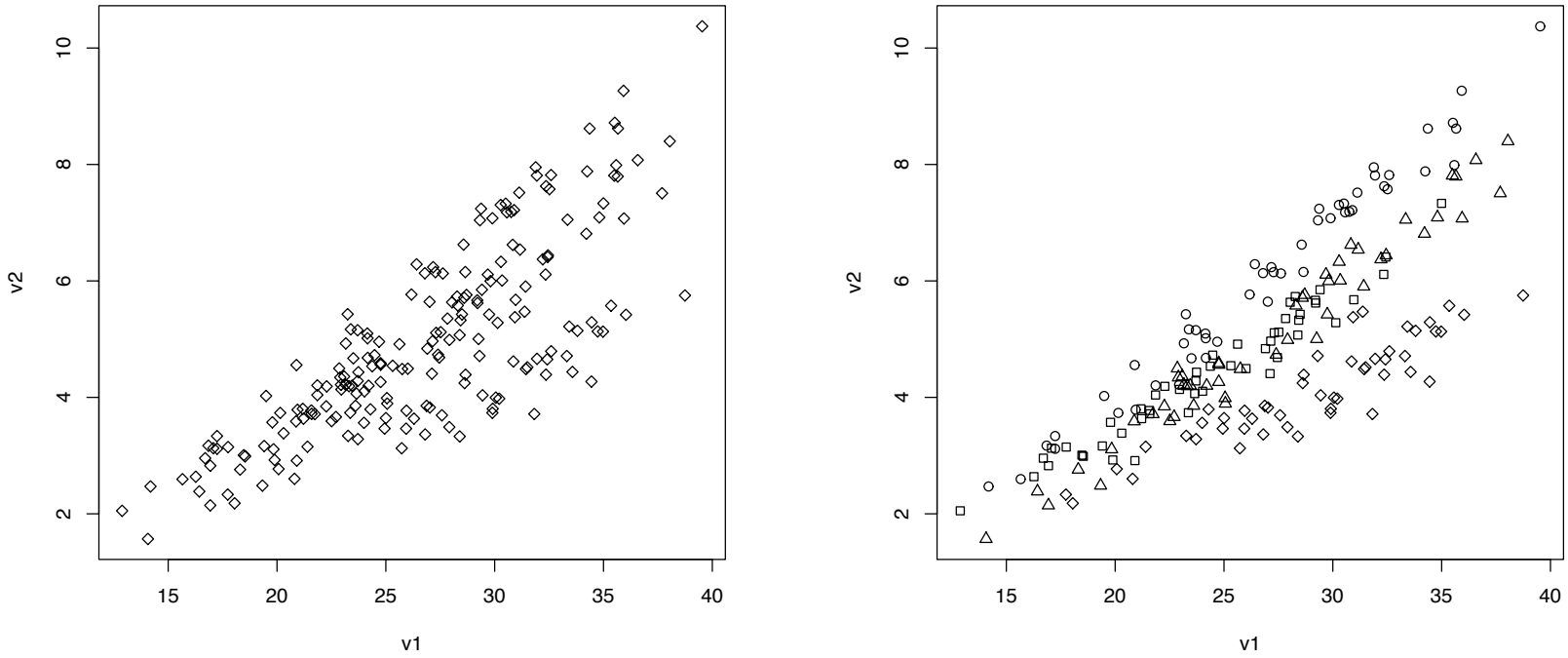


Figure: The left panel shows a view from GGobi. The right panel is the same view but with the cones labeled.

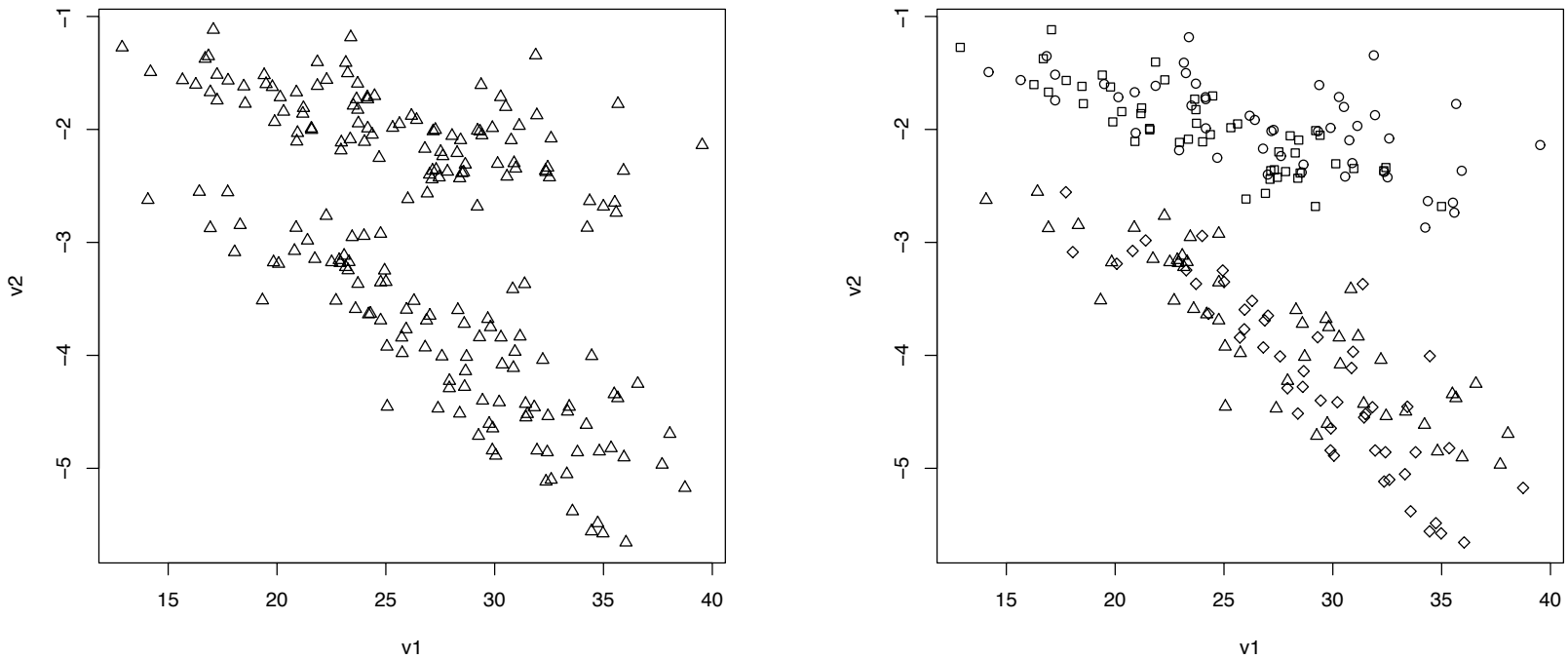


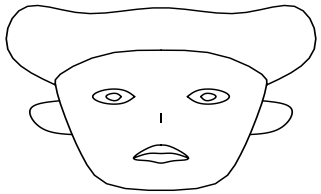
Figure: The left panels shows a view found from GGobi. The right panels is the same view but with the cones labeled.

Chernoff Faces

- Chernoff (1973) recognized that people are exquisitely sensitive to small differences in faces and proposed that this be harnessed to visualize high-dimensional data.
- Thus, under a mapping, a p-dimensional data point is converted to a list of values that specify features of a human face. For instance, the values of $x_{1,i}$ may represent the height of a face, the values of $x_{2,i}$ might represent the width of a face, and so forth. Then, each face generated from the data points has a unique expression.
- Using the Canadian household expenditure data for which stars and profiles were plotted before, gives the Chernoff faces in the next figure.

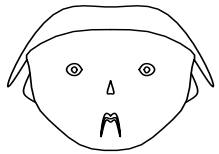
Canadian Household Expenditures

Ottawa

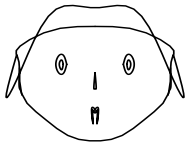


Toronto

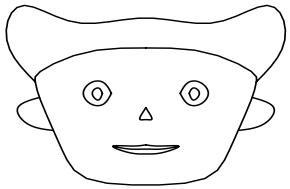
Montreal



Quebec



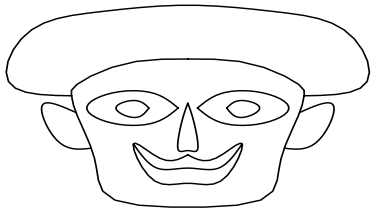
Vancouver



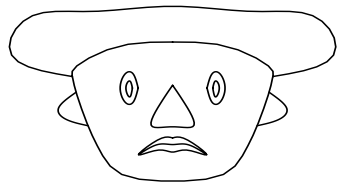
Edmonton

Saskatoon

Calgary



Victoria



- Remember: A Hilbert space $\mathcal{H}$ is a complete normed linear space where the norm comes from an inner product.
- Riesz Representation Theorem: Every continuous linear functional L on $\mathcal{H}$ has a unique kenel g_L so that $L(f) = \langle g_L, f \rangle$.
- Let $[x]$ be the evaluation functional, $[x](f) = f(x)$.
- RRT implies $\exists K_x$ so that $\langle K_x, f \rangle = f(x)$.
- The symmetric function $K(x, y) = K_x(y) = \langle K_x, K_x \rangle$ is the reproducing kernel of $\mathcal{H}$ because

$$\forall x \quad \langle K(x, \cdot), f(\cdot) \rangle = f(x).$$

- You can start with a Hilbert space and find its reproducing kernel or you can start with a reproducing kernel and construct its Hilbert space. K is always symmetric, positive definite, and reproducing.

- There is a huge RKHS-based approach to function approximation. One major result:
- **Representer Theorem:** Let $\Omega : [0, \infty) \to I\!R$ be a strictly monotonic increasing function, $\mathcal{X}$ be a set, and $c : (\mathcal{X} \times I\!R^2)^n \to I\!R \cup \{\infty\}$ be an arbitrary loss function. Then each minimizer $f \in \mathcal{H}$ of the regularized risk functional

$$c((x_1, y_1, f(x_1)), \cdots, (x_n, y_n, f(x_n))) + \Omega(\|f\|_{\mathcal{H}})$$

admits a representation of the form

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x).$$

- The form of the solution in an RKHS is determined up to n constants, independent of p.

- There are many applications of the Representer Theorem in classification, regression, dimension reduction etc etc.
- Let's look at the Relevance Vector Machine (RVM) of Tipping (2001).
- To present the RVM, choose a kernel and write the 'model'

$$Y = \sum_{j=1}^{n} w_j K(x, x_j) + \epsilon,$$

in which w_0 has been set to zero.

- This is not a model in the strict classical sense.
- Want $w = (w_1, \cdots, w_k)$ & $h(x) = (K(x, x_1^*), \ldots, K(x, x_k^*))'$ wth $k << n$ so that we can 'fit'

$$Y = w^{\perp} h(x) + \epsilon$$

- Many ways to estimate w and hence choose which vectors x_i are 'relevant' i.e., appear in the solution.
- First, a regularization approach is possible, as in LASSO for instance. More later.
- A simpler approach is truncation using a posterior threshold in a Bayesian context.
- That is, put a Normal prior on w to induce closed-form expressions for almost all the important estimation and prediction equations.
- To specify the RVM regression, write

$$Y = Hw + \epsilon,$$

where $y = (y_1, \cdots, y_n)^T$, $w = (w_1, \cdots, w_n)^T$, and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T$.

- The design matrix is

$$H = \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \vdots & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{bmatrix}.$$

- If the error term is $\epsilon_i \overset{IID}{\sim} N(0, \sigma^2)$ then the likelihood function for an IID sample is

$$p(y|H, w, \sigma^2) = N(y|Hw, \sigma^2 I_n).$$

- The parameter vector $\theta = (w, \sigma^2)$ is $(n+1)$-dimensional, and there are n data points for estimating it. This leads to non-unique solutions. This problem disappears (technically) if the variance σ^2 is known, but this is unrealistic in practice.

- An alternative Bayesian solution uses independent zero-mean normal priors for the coefficients in w_i. Set

$$p(w_i|\alpha_i) = N(w_i|0, \alpha_i^{-1}),$$

  so that $p(w|\alpha) = N_n(\mathbf{0}, \mathbf{D})$, in which $\mathbf{D} = \text{Diag}(\alpha_1^{-1}, \ldots, \alpha_n^{-1})$ and α denotes the vector $(\alpha_1, \ldots, \alpha_n)^T$.
- Normal priors don't usually give sparsity. However, using a Gamma hyperprior on each α_i yields a Student-t marginal for for w_i when α_i is integrated out, and this leads to sparsity of a sort.
- Even though each individual Student t for w_i is no candidate for sparsity, their product $p(w) = \prod_{i=1}^n p(w_i)$ has a surface that induces a sparsity pressure on w.

- That is, choose

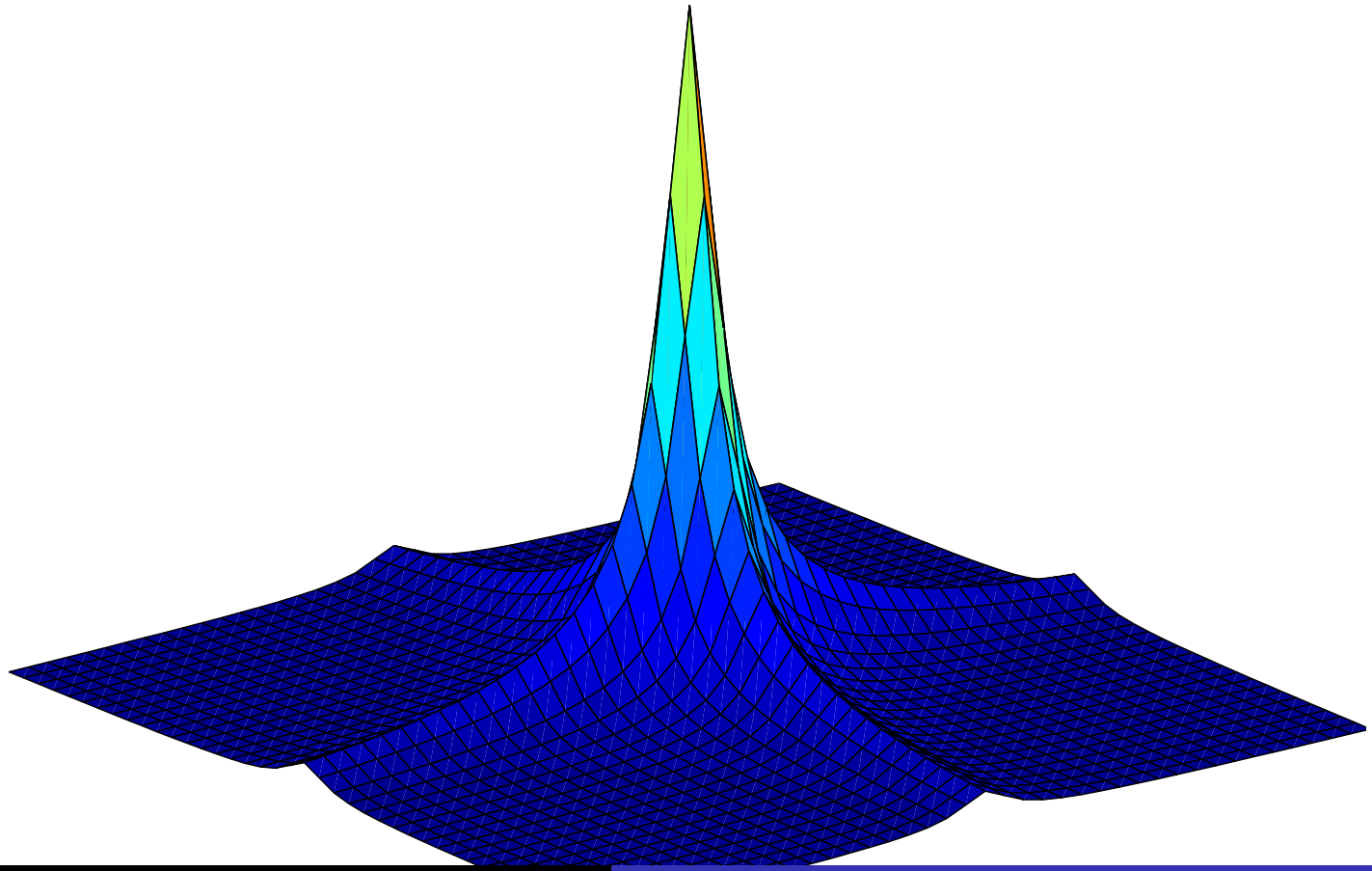
$$p(\alpha_i|a,b) = \text{Gamma}(\alpha_i|a,b).$$

- The marginal prior density for w_i is

$$\begin{aligned} p(w_i) &= \int p(w_i|\alpha_i)p(\alpha_i)d\alpha_i \\ &= \frac{b^a\Gamma(a+1/2)}{(2\pi)^{1/2}\Gamma(a)}(b+w_i^2/2)^{-(a+(1/2))}. \end{aligned}$$

- So, the joint prior for the w is a product of independent Student-t distributions over the w_i's.
- This prior (surprisingly) exhibits sparsity.

Two Dimensional Case



- Now, relevant vectors can be obtained. Suppose a weight w_i has variance α_i^{-1} tending to zero. Then, the distribution of w_i is sharply peaked at zero, and the corresponding vector x_i is irrelevant. All the vectors for which the variance α_i^{-1} does not tend to zero are relevant.
- In practice, the RVM is easily determined by truncation: Choose a large threshold for α_i, and set α_i^{-1} to zero if α_i is greater than the threshold.
- That is, relevant vectors are those for which the data do not permit the distribution of α_i to be too concentrated at zero.
- It remains to be seen how the prior combines with the likelihood to give a sparse posterior density.

- Assume σ^2 is known. The posterior is

$$p(w, \alpha|y) \propto p(y|H, w, \sigma^2)p(w|\alpha)p(\alpha|a, b)$$

where the Bayes model under normal noise with variance σ^2 is

$$p(y|H, w, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\|y-Hw\|^2},$$

$$p(w|\alpha) = \prod_{i=1}^{n} p(w_i|\alpha_i) \qquad \text{and} \qquad p(\alpha|a, b) = \prod_{i=1}^{n} p(\alpha_i|a, b).$$

- The marginal posterior $p(w|y)$ is obtained from $p(w, \alpha|y)$ by integration and specification of a and b.
- Want $p(w|y)$ to have the same form as in $p(w)$ in the figure. That is, want the effect of the prior specification to favor values of α along the axes or at the origin.

- Even with σ^2 known, the marginal posteriors,

$$p(w|y) = \int p(w, \alpha|y) d\alpha = \frac{p(y|w)p(w)}{p(y)}$$

and

$$p(\alpha|y) = \int p(w, \alpha|y) dw = \frac{p(y|\alpha)p(\alpha)}{p(y)},$$

cannot be computed in closed form.

- However, empirical Bayes approximation techniques can be used to obtain estimates of w and α. Alternatively, Markov chain Monte Carlo techniques can also be used to explore the joint posterior $p(w, \alpha|y)$.

- A sandard derivation gives that the conditional posterior density for w is

$$p(w|\alpha, \sigma^2, y) = N(w; \mu, V),$$

where

$$V = (H^T \sigma^2 I_n H + \mathbf{A})^{-1} \text{ and } \mu = V H^T \sigma^2 . I_n y.$$

- A more elaborate yet still standard derivation shows that the marginal likelihood $p(\mathbf{y}|\alpha, \sigma^2)$ is given by

$$p(y|\alpha, \sigma^2) = N(y; 0, \sigma^2 I_n + H A^{-1} H^T),$$

where $A = \text{Diag}(\alpha_1, \ldots, \alpha_n) = D^{-1}$.

- The two most important quantities, namely α and σ^2, are estimated by finding the values that maximize $\ln p(y|\alpha, \sigma^2)$.

- As shown in Tipping (2001), it turns out that finding

$$(\hat{\alpha}, \hat{\sigma}^2) = \arg \max_{\alpha, \sigma^2} \ln p(y|\alpha, \sigma^2)$$

reduces to a two-step iterative procedure: Initialize α and σ^2, and use them to obtain the posterior covariance matrix V and the posterior mean μ. Then, let μ_i be the ith component of μ and $\gamma_i = 1 - \alpha_i V_{ii}$.

- The iteration proceeds by setting

$$\alpha_i^{(\text{new})} = \frac{\gamma_i}{\mu_i^2}, \quad \text{and} \quad (\sigma^2)^{(\text{new})} = \frac{\|y - H\mu\|^2}{n - \sum_{i=1}^{n} \gamma_i},$$

and then recalculating until convergence.

- Using these prior specifications, RVM typically gives a regression function sparser than other methods and therefore often gives better predictive performance.