Stat 577 "Statistical Learning Theory"

T6: Dimension reduction, variational inference, Frechet regression

Xiongzhi Chen

Washington State University

Classic PCA

Population version

Suppose $\{x_i\}_{i=1}^N$ are i.i.d. observations of $X = \left(z_1, ..., z_p\right)^T \in \mathbb{R}^p$ with covariance matrix Σ. Then we have a formulation of the population version PCA in spectral decomposition:

- Spectral decomposition of Σ: $\Sigma = UDU^T$ with orthogonal matrix $U \in \mathbb{R}^{p \times p}$ and diagonal matrix $D = \text{diag}\left\{\delta_1, \ldots, \delta_p\right\} \in \mathbb{R}^{p \times p}$ such that
$$\delta_i \geq \delta_{i+1} > 0 \text{ for each } i$$
- p linear transforms: $U = \left(l_1, \cdots, l_p\right)$ (each l_i is a column vector) such that $y_i = \langle l_i, X \rangle = X^T l_i$, where we recall that $\langle l_i, X \rangle$ is the inner product of l_i and X

PCA finds p linear combinations of $\{z_i\}_{i=1}^p$, i.e., find p vectors $\gamma_i = \left(\gamma_{i1}, ..., \gamma_{ip}\right)^T$ and form $g_i = \sum_{j=1}^p \gamma_{ij} z_j = \langle \gamma_i, X \rangle$ such that the variances of $\{g_i\}_{i=1}^p$ are successively maximized and that $\{\gamma_i\}_{i=1}^p$ are orthogonal

Population version

Proof of the claim in the last slide:

- Recall $\Sigma = UDU^T$ and $y_i = X^T l_i$ with $U = (l_1, \cdots, l_p)$ and $D = \text{diag}\{\delta_1, \ldots, \delta_p\}$
- Fact 1: y_i and y_j for $i \neq j$ are uncorrelated, i.e.,

$$\text{Cov}(y_i, y_j) = l_i^T \text{Cov}\left(X, X^T\right) l_j = l_i^T \Sigma l_j = \delta_j l_i^T l_j = 0$$

- Fact 2: $l_1 = \text{argmax}_{\|z\|=1} \text{Var}\left(X^T z\right)$ and $\max_{\|z\|=1} z^T \Sigma z = \delta_1$, and for $2 \leq i \leq p$

$$l_i = \text{argmax}\left\{\text{Var}\left(X^T z\right) : \|z\| = 1, \langle z, l_j \rangle = 0, \forall\, 1 \leq j \leq i-1\right\}$$

and the maximum in the above is δ_i, since

$$\text{Var}\left(Y^T z\right) = z^T S z \leq \|z\|^2 \lambda_{\max}(S)$$

for any $Y \in \mathbb{R}^p$ with covariance matrix S, where $\lambda_{\max}$ is the largest eigenvalue of S, and $U^T U = I$

Data version

- Let $X \in \mathbb{R}^n$ and $\mathbf{X} = \left(\boldsymbol{x}_1^T, ..., \boldsymbol{x}_m^T\right)^T$ be the *column-centered* data matrix (i.e., sample mean for entries of each column of $\mathbf{X}$ is 0)
- SVD $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ gives $\mathbf{S} = \mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$, and $\mathbf{S} = m^{-1}\widetilde{\text{cov}}\,(X)$
- Let $\left\{\left(d_i^2, \mathbf{v}_i\right)\right\}_{i=1}^n$ be the eigenpairs of $\mathbf{S}$. Given any orthonormal $\{\tilde{\mathbf{u}}_i\}_{i=1}^n$, define $\mathbf{z}_i = \mathbf{X}\tilde{\mathbf{u}}_i$. Then

$$\mathbf{z}_i = (z_{1i}, ..., z_{mi})^T = \left(\langle \boldsymbol{x}_1, \tilde{\mathbf{u}}_i \rangle, \cdots, \langle \boldsymbol{x}_m, \tilde{\mathbf{u}}_i \rangle\right)^T$$

  and $\widetilde{\text{cov}}\,(\mathbf{z}_i) = \tilde{\mathbf{u}}_i^T\mathbf{S}\tilde{\mathbf{u}}_i$. (What is the interpretation of $\mathbf{z}_i$?)
- SMS Lemma (applied to $\mathbf{S}$) implies that $\{\mathbf{v}_i\}_{i=1}^n$ are the *directions* for which $\{\mathbf{z}_i\}_{i=1}^n$ *successively* achieve *maximal sample variances*, i.e.,

$$\begin{cases} \max_{\|\mathbf{v}\|=1, \mathbf{v} \perp \langle \mathbf{v}_1, ..., \mathbf{v}_k \rangle} \widetilde{\text{var}}\left(\langle \mathbf{X}, \mathbf{v} \rangle\right) = m^{-1} d_{k+1}^2 \\ \text{argmax}_{\|\mathbf{v}\|=1, \mathbf{v} \perp \langle \mathbf{v}_1, ..., \mathbf{v}_k \rangle} \widetilde{\text{var}}\left(\langle \mathbf{X}, \mathbf{v} \rangle\right) = \mathbf{v}_{k+1} \end{cases}$$

- Thus, optimal $\mathbf{z}_i = d_i \mathbf{v}_i$

Recap

- Population version of optimal projections:
  - Let $X \in \mathbb{R}^n$ have covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$ with eigenpairs $(\sigma_i, \mathbf{v}_i)$
  - Optimal linear combinations of entries of X are $\{y_i = \langle X, \mathbf{v}_i \rangle\}_{i=1}^n$ when $\{\mathbf{v}_i\}_{i=1}^n$ have to be orthonormal
  - $\{y_i\}_{i=1}^n$ *successively* achieve *maximal variances and are mutually uncorrelated*
- Sample version of optimal projections:
  - Let $X \in \mathbb{R}^n$ and $\mathbf{X} = \left(\boldsymbol{x}_1^T, ..., \boldsymbol{x}_m^T\right)^T$ be the *column-centered* data matrix
  - Let $\left\{\left(d_i^2, \mathbf{v}_i\right)\right\}_{i=1}^n$ be the eigenpairs of $\mathbf{S} = \mathbf{X}^T\mathbf{X}$
  - Optimal linear combinations of columns of $\mathbf{X}$ are $\{\mathbf{z}_i = \mathbf{X}\mathbf{v}_i\}_{i=1}^n$ when $\{\mathbf{v}_i\}_{i=1}^n$ have to be orthonormal
  - $\{\mathbf{z}_i\}_{i=1}^n$ *successively* achieve *maximal sample variances*
- Both versions involve orthogonal projections

PCA via optimal subspace

Data version and SVD

Formulation Two:

- PCA provides the best linear approximate to $\mathbf{X}$ under the Frobenius norm among all subspace of dimension $q \leq p$.

In order to understand this, we need "singular value decomposition (SVD)". The SVD of the centered data matrix $\mathbf{X}_{N\times p}$ is

$$\mathbf{X} = \mathbf{U}_{N\times p}\mathbf{D}_{p\times p}\mathbf{V}_{p\times p}^T = \sum_{i=1}^{\mathsf{rank}(\mathbf{X})} d_i \mathbf{u}_i \mathbf{v}_i^T \text{ with } \mathbf{U}^T\mathbf{U} = \mathbf{I}_p = \mathbf{V}^T\mathbf{V}$$

- rank denotes the rank of a matrix; $p = \mathsf{rank}\,(\mathbf{X})$; $\mathbf{I}_s$ is the identity matrix of dimension s
- $\mathbf{D}$ is a diagonal matrix such that $\mathbf{D} = \mathsf{diag}\left\{d_1, \ldots, d_p\right\}$ and $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$; each d_i is called a "singular value"
- $\mathbf{u}_i$ is the ith column of $\mathbf{U}$ and is called a "left singular vector", and $\mathbf{v}_i$ the ith column of $\mathbf{V}$ and is called a "right singular vector"

Data version and SVD

- Recall the SVD

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \sum_{i=1}^{\mathsf{rank}(\mathbf{X})} d_i \mathbf{u}_i \mathbf{v}_i^T \text{ with } \mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$$

- If we compare SVD with the spectral decomposition of $\mathbf{X}^T\mathbf{X}$, we see that

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$$

- So, each eigenvalue λ_i of $\mathbf{X}^T\mathbf{X}$ satisfies $\lambda_i = d_i^2$ and the eigenvector $\mathbf{a}_i$ of $\mathbf{X}^T\mathbf{X}$ can be identified with $\mathbf{v}_i$.

Fomulation via regression

- Given N observations $\{x_i\}_{i=1}^N$, consider the rank q linear model

$$f(\lambda) = \mu + \mathbf{V}_q \lambda$$

for representing $\{x_i\}_{i=1}^N$, where the $p \times q$ matrix $\mathbf{V}_q$ has q orthonormal columns, $\mu \in \mathbb{R}^p$ is a location vector, and $\lambda \in \mathbb{R}^q$ contains parameters. Namely, we want to find λ_i such that $f(\lambda_i)$ reconstructs x_i for each i

- A least squares fit leads to the optimization problem

$$\min_{\mu, \{\lambda_i\}, \mathbf{V}_q} \sum_{i=1}^N \left\| x_i - \mu - \mathbf{V}_q \lambda_i \right\|^2$$

whose critical point satisfies $\hat{\mu} = \bar{x}$ and $\hat{\lambda}_i = \mathbf{V}_q^T (x_i - \bar{x})$

Fomulation via regression

- Now we only need to solve

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} \left\| (x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x}) \right\|^2$$

- Since $\mathbf{X}$ is already centered, i.e., $\bar{x} = 0$ holds, the above reduces to

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} \left\| x_i - \mathbf{V}_q \mathbf{V}_q^T x_i \right\|^2 = \min_{\mathbf{V}_q} \left\| \mathbf{X}^T - \mathbf{V}_q \mathbf{V}_q^T \mathbf{X}^T \right\|_F^2,$$

where for any matrix A, its Frobenius norm is defined as

$$\|A\|_F = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{p'} a_{ij}^2} = \text{trace}\left(A^T A\right) = \text{trace}\left(A A^T\right) = \sum_{l=1}^{\min\{p,p'\}} d_l^2,$$

where d_i are the singular values of A and trace denotes the trace of a square matrix

Residual sum of squares

- The solution of

$$\min_{\mathbf{V}_q} \left\| \mathbf{X}^T - \mathbf{V}_q \mathbf{V}_q^T \mathbf{X}^T \right\|_F^2$$

  is $\hat{\mathbf{V}}_q = (\mathbf{v}_1, ..., \mathbf{v}_q)$ for which $\mathbf{v}_i$ is the ith column of $\mathbf{V}$ (and hence is an eigenvector associated with the ith largest eigenvalue λ_i of $\mathbf{X}^T\mathbf{X}$)
- This can be seen from the SVD

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \sum_{i=1}^{\mathsf{rank}(\mathbf{X})} d_i \mathbf{u}_i \mathbf{v}_i \text{ with } \mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$$

  since

$$\begin{aligned} \left\| \mathbf{X}^T - \mathbf{V}_q \mathbf{V}_q^T \mathbf{X}^T \right\|_F^2 &= \text{trace}\left(\left(\mathbf{I} - \mathbf{V}_q \mathbf{V}_q^T \right) \mathbf{V}\mathbf{D}^2\mathbf{V}^T \left(\mathbf{I} - \mathbf{V}_q \mathbf{V}_q^T \right) \right) \\ &= \text{trace}\left(\mathbf{D}^2\mathbf{V}^T \left(\mathbf{I} - \mathbf{V}_q \mathbf{V}_q^T \right) \mathbf{V} \right) \end{aligned}$$

Residual sum of squares

Set $\delta = \text{trace}\left(\mathbf{D}^2\left(\mathbf{I} - \left(\mathbf{V}_q^T\mathbf{V}\right)^T\mathbf{V}_q^T\mathbf{V}\right)\right)$. Then $\left\|\mathbf{X}^T - \mathbf{V}_q\mathbf{V}_q^T\mathbf{X}^T\right\|_F^2 = \delta$.

- Since $\mathbf{V}_q$ contains q orthonormal vectors, we have

$$\mathbf{I} - \mathbf{V}_q\mathbf{V}_q^T = \left(\mathbf{I} - \mathbf{V}_q\mathbf{V}_q^T\right)\left(\mathbf{I} - \mathbf{V}_q\mathbf{V}_q^T\right),$$

  i.e., $\mathbf{I} - \mathbf{V}_q\mathbf{V}_q^T$ is idempotent and an orthogonal projection.

- Thus, $\mathbf{Q} = \mathbf{I} - \left(\mathbf{V}_q^T\mathbf{V}\right)^T\mathbf{V}_q^T\mathbf{V}$ is idempotent and an orthogonal projection, and can only have eigenvalues 0 or 1, and

$$\text{trace}\left(\mathbf{Q}\right) = \text{rank}\left(\mathbf{Q}\right) = p - q.$$

  Further, there exits an orthogonal matrix $\mathbf{K}$ such that

$$\mathbf{K}\mathbf{Q}\mathbf{K}^T = \text{diag}\left\{\mathbf{I}_{n-q}, \mathbf{0}_q\right\} \text{ or } \mathbf{K}\mathbf{Q}\mathbf{K}^T = \text{diag}\left\{\mathbf{0}_q, \mathbf{I}_{n-q}\right\}.$$

Residual sum of squares

- Let s_i be the ith entry of $\text{diag}\left\{\mathbf{I}_{p-q}, \mathbf{0}_q\right\}$ or $\text{diag}\left\{\mathbf{I}_{p-q}, \mathbf{0}_q\right\}$. Then

$$\begin{aligned}\delta &= \text{trace}\left(\mathbf{D}^2\mathbf{K}^T\text{diag}\left\{s_1, \ldots, s_p\right\}\mathbf{K}\right) \\ &= \text{trace}\left(\text{diag}\left\{s_1, \ldots, s_p\right\}\mathbf{K}\mathbf{D}^2\mathbf{K}^T\right)\end{aligned}$$

- Let $\mathbf{k}_i$ be the ith column of $\mathbf{K}$. Then $\text{trace}\left(\mathbf{k}_i\mathbf{k}_i^T\right) = 1$ for all i and

$$\delta = \sum_{i=1}^{p} d_i^2 s_i \text{trace}\left(\mathbf{k}_i\mathbf{k}_i^T\right) = \sum_{i=1}^{p} d_i^2 s_i$$

and the minimum of δ is $\sum_{i=q+1}^{p} d_i^2$, achieved by $\hat{\mathbf{V}}_q$, the $p \times q$ matrix that contains the first q columns of $\mathbf{V}$.

Solution as optimal subspace

Let $\langle \mathbf{V}_q \rangle$ be the linear space spanned by the columns of $\mathbf{V}_q$. Then $\mathbf{V}_q \mathbf{V}_q^T \mathbf{X}^T$ projects each row x_i of $\mathbf{X}$ onto $\langle \mathbf{V}_q \rangle$, and the Frobenius norm

$$\left\| \mathbf{X}^T - \mathbf{V}_q \mathbf{V}_q^T \mathbf{X}^T \right\|_F^2 = \left\| \left(\mathbf{I} - \mathbf{V}_q \mathbf{V}_q^T \right) \mathbf{X}^T \right\|_F^2$$

is the sum of residual squared Euclidean lengths of projecting $\{x_i\}_{i=1}^N$.

- Recall the SVD of $\mathbf{X}$ as $\mathbf{X} = \mathbf{U}_{N \times p} \mathbf{D}_{p \times p} \mathbf{V}_{p \times p}^T$. PCA provides the best linear approximate to $\mathbf{X}$ in Frobenius norm among all subspace of dimension $q \leq p$, and the optimal is achieved by the column space of $\hat{\mathbf{V}}_q$, the $p \times q$ matrix that contains the first q columns of $\mathbf{V}$ (associated with the q largest singular values of $\mathbf{X}$).
- The optimal $\hat{\lambda}_i = \hat{\mathbf{V}}_q^T x_i$, i.e., $\hat{\lambda}_i$ is the ith row of $\mathbf{UD}$ for $1 \leq i \leq q$. The columns of $\mathbf{UD}$ are called "principal components of $\mathbf{X}$".
- The fitted model is $x_i \approx \bar{x} + \hat{\mathbf{V}}_q \hat{\mathbf{V}}_q^T (x_i - \bar{x})$

Application

PCA: illustration (Figure 14.21 of Text)

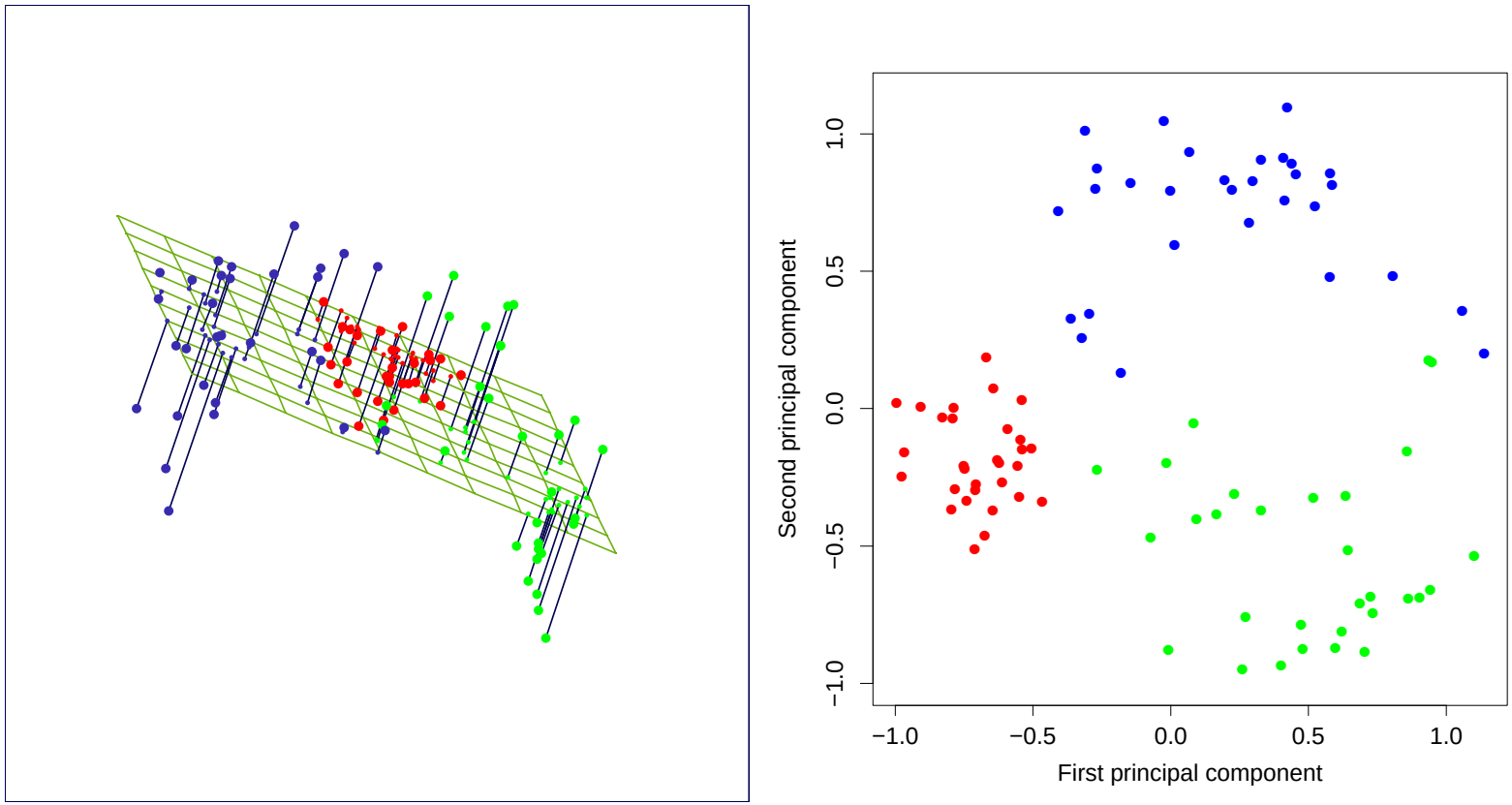


FIGURE 14.21. *The best rank-two linear approximation to the half-sphere data. The right panel shows the projected points with coordinates given by* $\mathbf{U}_2\mathbf{D}_2$*, the first two principal components of the data.*

PCA: example in Section 14.5 of Text

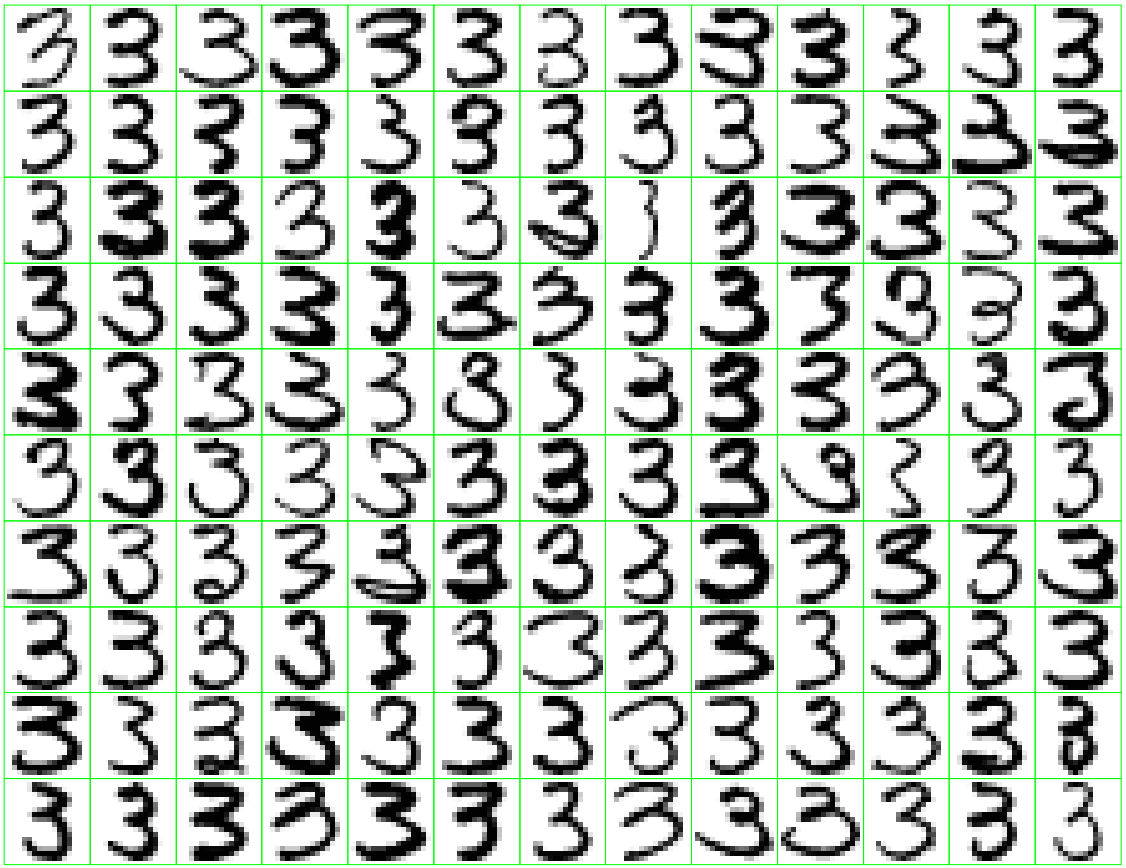


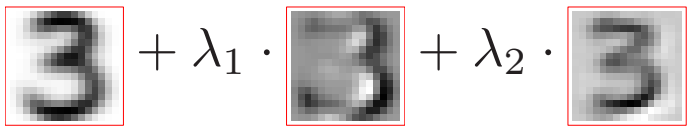
FIGURE 14.22. *A sample of* 130 *handwritten 3's shows a variety of writing styles.*

The ith handwritten 3 is represented as a 16×16 grayscale image and then vectorized as $x_i \in \mathbb{R}^{256}$ for $1 \leq i \leq 130$

PCA: example

Take $v_1, v_2 \in \mathbb{R}^{256}$ and $\lambda = (\lambda_1, \lambda_2)^T \in \mathbb{R}^2$, where v_1 encodes horizontal movement and v_2 vertical movement. We get the model

$$\begin{aligned} \hat{f}(\lambda) &= \bar{x} + \lambda_1 v_1 + \lambda_2 v_2 \\ &= \end{aligned}$$



(14.55)

that approximates the data matrix $X \in \mathbb{R}^{160 \times 256}$ by the 2-dimensional, column space of $\mathbf{V}_2 = (v_1, v_2)$ under the Frobenius norm

PCA: example

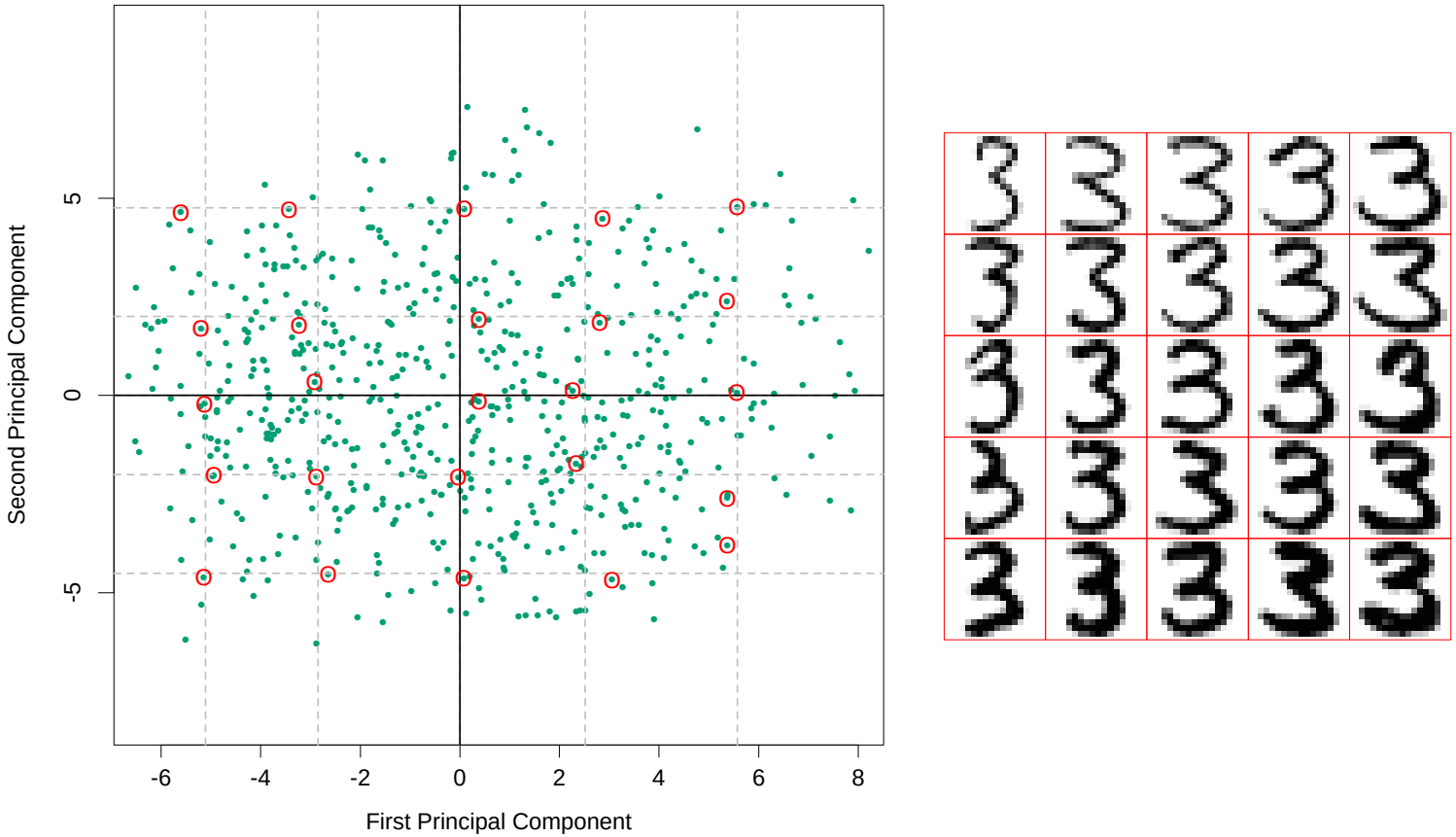


FIGURE 14.23. *(Left panel:) the first two principal components of the hand-written threes. The circled points are the closest projected images to the vertices of a grid, defined by the marginal quantiles of the principal components. (Right panel:) The images corresponding to the circled points. These show the nature of the first two principal components.*

PCA: example

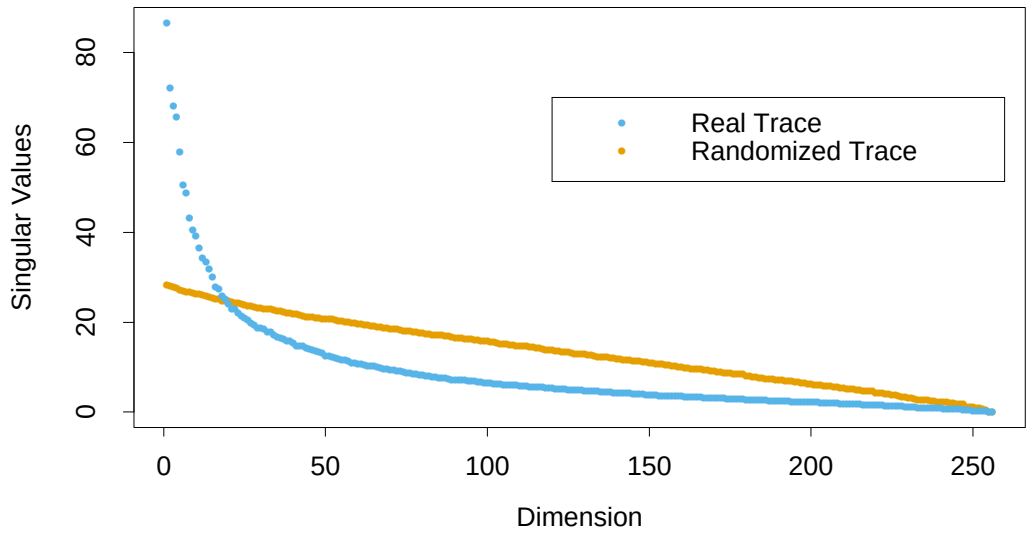


FIGURE 14.24. *The 256 singular values for the digitized threes, compared to those for a randomized version of the data (each column of* $\mathbf{X}$ *was scrambled).*

Among the 256 singular values of X, approximately 50 account for 90% of the variation in the 3's, and 12 account for 63%. For this example, a relatively small subset of the principal components serve as excellent lower-dimensional features for representing the high-dimensional data.

Sparse PCA

Motivation behind sparse PCA

Recall $\mathbf{X} = \mathbf{UDV}^T$ via the SVD

- Columns of $\mathbf{UD}$ are the principal components (PCs)
- Columns of $\mathbf{V}$ are the corresponding loadings of the principal components

For usual PCA, the PC's are hard to interpret since:

- Each PC is a linear combination of all p variables
- The loadings are typically nonzero

Sparse PCA aims to produce PCs with sparse loadings, i.e., most loadings are zero. To illustrate this, we will use the paper "Sparse Principal Component Analysis" by Hui Zou, Trevor Hastie and Robert Tibshirani.

Sparse PCA: data version

- $\mathbf{X}_{N\times p}$ is the centered data matrix; X_i the ith row of $\mathbf{X}$
- Ridge regression:

$$(\hat{\alpha},\hat{\beta}) = \underset{\alpha\in\mathbb{R}^{p\times k},\beta\in\mathbb{R}^{p\times k};\alpha^T\alpha=I}{\operatorname{argmin}} \sum_{i=1}^{N}\left\|X_i-\alpha\beta^T X_i\right\|^2+\lambda\sum_{j=1}^{k}\|\beta_j\|^2 \qquad (1)$$

  then $\hat{\beta}_i \propto \mathbf{V}_i$, where β_j is the jth column of β, $\propto$ denotes "proportional to" and λ is the tuning parameter
- Remark: (1) can be written as

$$(\hat{\alpha},\hat{\beta}) = \underset{\alpha\in\mathbb{R}^{p\times k},\beta\in\mathbb{R}^{p\times k};\alpha^T\alpha=I}{\operatorname{argmin}} \left\|X^T-\alpha\beta^T X^T\right\|_F^2+\lambda\|\beta\|_F^2$$

- Recall PCA as

$$\underset{\mathbf{V}_q\in\mathbb{R}^{p\times q},\operatorname{rank}\left(\mathbf{V}_q\right)=q,\mathbf{V}_q^T\mathbf{V}_q=I}{\operatorname{argmin}} \left\|X_i-\mathbf{V}_q\mathbf{V}_q^T X_i\right\|_F^2$$

Sparse PCA: data version

- $\mathbf{X}_{N\times p}$ is the centered data matrix; X_i the ith row of $\mathbf{X}$
- Ridge and LASSO:

$$(\hat{\alpha},\hat{\beta}) = \underset{\alpha\in\mathbb{R}^{p\times k},\beta\in\mathbb{R}^{p\times k};\alpha^T\alpha=I}{\operatorname{argmin}} \left\|X^T - \alpha\beta^T X^T\right\|_F^2 + \lambda\left\|\beta\right\|_F^2 + \sum_{j=1}^{k}\lambda_{1,j}\left\|\beta_j\right\|_1 \tag{2}$$

  then $\hat{\beta}_i$ is a sparse loading vector, where the l_1-norm $\|\cdot\|_1$ for $x=(x_1,...,x_s)\in\mathbb{R}^s$ is defined as $\|x\|_1=\sum_{i=1}^{s}|x_i|$, and $\left\{\lambda_{1,j}\right\}_{j=1}^{k}$ are tuning parameters
- The l_1-norm $\left\|\beta_j\right\|_1$ as a penalty on the β_j's will force entries of a β_j to be 0, leading to the so called "sparse loadings" and "sparse PCA". The strategy (2) applies to high-dimensional data (where $N<p$)
- **Caution**: unlike PCA, the principal components from sparse PCA are not necessarily uncorrelated

Sparse PCA: illustration

- The data set contain $n = 180$ observations and $p = 13$ measured variables (i.e., 13 features)
- This data set is the classic example showing the difficulty of interpreting principal components.
- The first 6 PCs will be examined
- In each of the tables to be given in the next few slides, the "Variance (%)" is the percentage of variance explained by a PC and "Cumulative Variance (%)" the percentage of variance explained together by the corresponding PC's

Sparse PCA: illustration

Table 1: *Pitprops data: loadings of the first 6 principal components*

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.404	0.218	-0.207	0.091	-0.083	0.120
length	-0.406	0.186	-0.235	0.103	-0.113	0.163
moist	-0.124	0.541	0.141	-0.078	0.350	-0.276
testsg	-0.173	0.456	0.352	-0.055	0.356	-0.054
ovensg	-0.057	-0.170	0.481	-0.049	0.176	0.626
ringtop	-0.284	-0.014	0.475	0.063	-0.316	0.052
ringbut	-0.400	-0.190	0.253	0.065	-0.215	0.003
bowmax	-0.294	-0.189	-0.243	-0.286	0.185	-0.055
bowdist	-0.357	0.017	-0.208	-0.097	-0.106	0.034
whorls	-0.379	-0.248	-0.119	0.205	0.156	-0.173
clear	0.011	0.205	-0.070	-0.804	-0.343	0.175
knots	0.115	0.343	0.092	0.301	-0.600	-0.170
diaknot	0.113	0.309	-0.326	0.303	0.080	0.626
Variance (%)	32.4	18.3	14.4	8.5	7.0	6.3
Cumulative Variance (%)	32.4	50.7	65.1	73.6	80.6	86.9

Sparse PCA: illustration

Table 2: *Pitprops data: loadings of the first 6 modified PCs by SCoTLASS*

$t = 1.75$						
Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	0.546	0.047	-0.087	0.066	-0.046	0.000
length	0.568	0.000	-0.076	0.117	-0.081	0.000
moist	0.000	0.641	-0.187	-0.127	0.009	0.017
testsg	0.000	0.641	0.000	-0.139	0.000	0.000
ovensg	0.000	0.000	0.457	0.000	-0.614	-0.562
ringtop	0.000	0.356	0.348	0.000	0.000	-0.045
ringbut	0.279	0.000	0.325	0.000	0.000	0.000
bowmax	0.132	-0.007	0.000	-0.589	0.000	0.000
bowdist	0.376	0.000	0.000	0.000	0.000	0.065
whorls	0.376	-0.065	0.000	-0.067	0.189	-0.065
clear	0.000	0.000	0.000	0.000	-0.659	0.725
knots	0.000	0.206	0.000	0.771	0.040	0.003
diaknot	0.000	0.000	-0.718	0.013	-0.379	-0.384
Number of nonzero loadings	6	7	7	8	8	8
Variance (%)	27.2	16.4	14.8	9.4	7.1	7.9
Adjusted Variance (%)	27.2	15.3	14.4	7.1	6.7	7.5
Cumulative Adjusted Variance (%)	27.2	42.5	56.9	64.0	70.7	78.2

Sparse PCA: illustration

Table 3: *Pitprops data: loadings of the first 6 sparse PCs by SPCA*

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.477	0.000	0.000	0	0	0
length	-0.476	0.000	0.000	0	0	0
moist	0.000	0.785	0.000	0	0	0
testsg	0.000	0.620	0.000	0	0	0
ovensg	0.177	0.000	0.640	0	0	0
ringtop	0.000	0.000	0.589	0	0	0
ringbut	-0.250	0.000	0.492	0	0	0
bowmax	-0.344	-0.021	0.000	0	0	0
bowdist	-0.416	0.000	0.000	0	0	0
whorls	-0.400	0.000	0.000	0	0	0
clear	0.000	0.000	0.000	-1	0	0
knots	0.000	0.013	0.000	0	-1	0
diaknot	0.000	0.000	-0.015	0	0	1
Number of nonzero loadings	7	4	4	1	1	1
Variance (%)	28.0	14.4	15.0	7.7	7.7	7.7
Adjusted Variance (%)	28.0	14.0	13.3	7.4	6.8	6.2
Cumulative Adjusted Variance (%)	28.0	42.0	55.3	62.7	69.5	75.8

Principal structures

Principal curve and surfaces

Principal curves and surfaces are generalizations of principal components, and they provide manifolds to approximate data. Let A be a non-empty (connected) set in $\mathbb{R}^m$, $\lambda = (\lambda_1, ..., \lambda_m)^T$ and

$$f(\lambda) = \left(f_1(\lambda), ..., f_p(\lambda)\right)^T \in \mathbb{R}^p$$

where each f_i is a scalar function. For each data value x of a random vector $X \in \mathbb{R}^p$, let $\lambda_f(x)$ be the closest point on the curve to x. Assume

$$f(\lambda) = E\left(X | \lambda_f(X) = \lambda\right)$$

- If $m = 1$, then f is called a "principal curve" for the distribution of X
- If $m = 2$, then f is called a "principal surface" for the distribution of X

Principal curve and surfaces

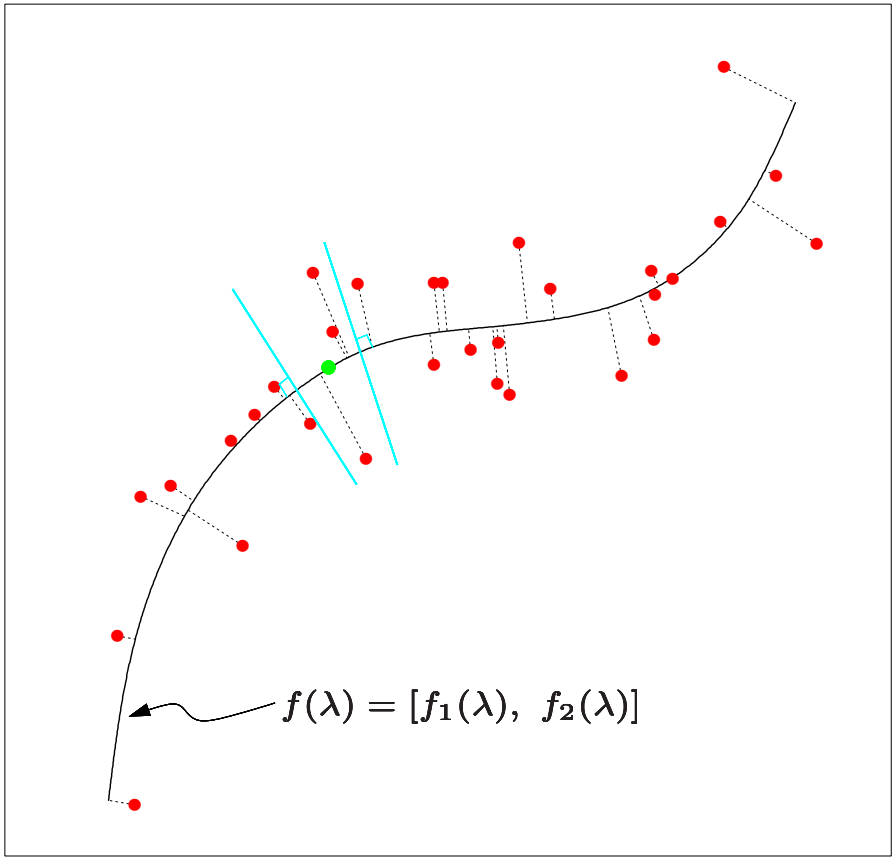


FIGURE 14.27. *The principal curve of a set of data. Each point on the curve is the average of all data points that project there.*

Principal curve and surfaces

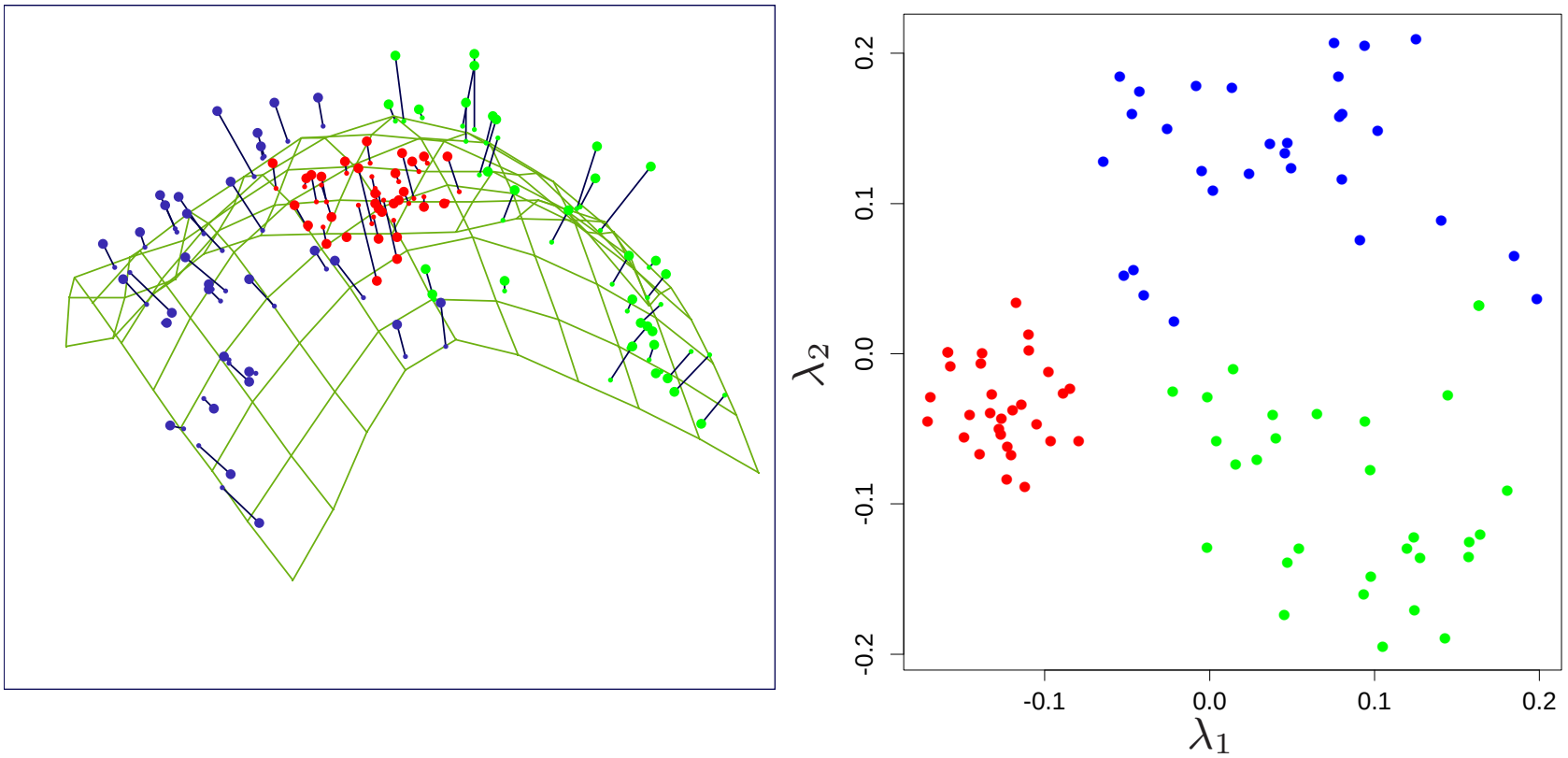


FIGURE 14.28. *Principal surface fit to half-sphere data. (Left panel:) fitted two-dimensional surface. (Right panel:) projections of data points onto the surface, resulting in coordinates* $\hat{\lambda}_1, \hat{\lambda}_2$.

Variational inference

Motivation

- To reduce computational complexity and achieve approximate inference without sacrificing too much accuracy
- Roughly 3 modern computational tools for analyzing massive data sets:
  - Convex relaxation such as the LASSO by Tibshirani (1996)
  - Variational inference, whose success can be contributed to the paper of Blei et al. (2003) and has connections with the "expectation-maximization (EM)" algorithm of Dempster et al. (1977)
  - Approximate Bayesian computation, originally proposed by Rubin (1984).

Bayseian model and Bayesian inference

Bayesian model and inference:

- the parameter $\theta \in \mathbb{R}^m$
- a prior $\pi(\theta)$ for θ; conditional density of data D given θ is $p(D|\theta)$
- Obtain posterior $p(\theta|D)$
- Conduct inference on θ using $p(\theta|D)$

Posterior

- Let $p(D)$ be marginal density of D
- Bayes rule implies

$$p(\theta|D) = \frac{\pi(\theta)\, p(D|\theta)}{p(D)}$$

- Joint distribution of (D, θ) as $p(D, \theta) = \pi(\theta)\, p(D|\theta)$
- Then posterior

$$p(D) = \int \pi(\theta)\, p(D|\theta)\, d\theta \tag{3}$$

(integral can be very complicated)

Complicated posterior

- Posterior: $p(D) = \int \pi(\theta)\, p(D|\theta)\, d\theta$
- Example: u a latent variable and

$$p(\theta|D) = \int l(\theta|D,u)\, s(u)\, du \tag{4}$$

- Example: θ is of high-dimension
- Variational inference approximates $p(D)$ by not computing the integral $\int \pi(\theta)\, p(D|\theta)\, d\theta$ exactly

Strategy of variational inference

- Replace D by Y. So, $p(\theta|y) = \frac{p(y,\theta)}{p(y)}$ with $p(y) = \int p(y,\theta)\, d\theta$
- Target: approximate $p(y)$ or $\int p(y,\theta)\, d\theta$. Strategy:
  - Pick $\mathcal{Q} \ni q(\theta)$, and solve optimization problem

$$\hat{q} \in \underset{q\in\mathcal{Q}}{\operatorname{argmax}} \exp\left(\int q(\theta) \log\left(\frac{p(y,\theta)}{q(\theta)}\right) d\theta\right) \tag{5}$$

  - Set $\hat{q}(\theta)$ as an approximate of $p(\theta|y)$, and approximate of $p(y)$ by

$$\begin{aligned} &\max_{q\in\mathcal{Q}} \exp\left(\int q(\theta) \log\left(\frac{p(y,\theta)}{q(\theta)}\right) d\theta\right) \\ &= \exp\left(\int \hat{q}(\theta) \log\left(\frac{p(y,\theta)}{\hat{q}(\theta)}\right) d\theta\right) \end{aligned} \tag{6}$$

Lower bound on marginal

- Pick density $q(\theta)$ for θ and notice $p(y) = \frac{p(y,\theta)}{p(\theta|y)}$. Then

$$\begin{aligned}
\log p(y) &= (\log p(y)) \int q(\theta)\, d\theta \qquad (7)\\
&= \int q(\theta) \log p(y)\, d\theta = \int q(\theta) \log \left(\frac{p(y,\theta)}{p(\theta|y)} \right) d\theta \\
&= \int q(\theta) \log \left(\frac{p(y,\theta)/q(\theta)}{p(\theta|y)/q(\theta)} \right) d\theta \\
&= \int q(\theta) \log \left(\frac{p(y,\theta)}{q(\theta)} \right) d\theta - \int q(\theta) \log \left(\frac{p(\theta|y)}{q(\theta)} \right) d\theta, \qquad (8)
\end{aligned}$$

- Concavity of log and Jensen's inequality:

$$\int q(\theta) \log \left(\frac{p(\theta|y)}{q(\theta)} \right) d\theta \leq 0 \qquad (9)$$

Lower bound on marginal

Lemma

For any density function $q(\theta)$ of θ, it holds that

$$p(y) \geq \underline{p}(y;q) = \exp\left(\int q(\theta) \log\left(\frac{p(y,\theta)}{q(\theta)}\right) d\theta\right). \tag{10}$$

- Now we move onto $\operatorname{argmax}_{q\in\mathcal{Q}} \underline{p}(y;q)$, i.e.,

$$\operatorname*{argmax}_{q\in\mathcal{Q}} L(q;p(\cdot|y)) \text{ with } L(q;p(\cdot|y)) = \int q(\theta) \log\left(\frac{p(\theta|y)}{q(\theta)}\right) d\theta \tag{11}$$

- Kullback–Leibler (KL) divergence:

$$KL(P,Q) = \int P(\theta) \log \frac{P(\theta)}{Q(\theta)} d\theta \text{ and } KL(P,Q) \geq 0$$

Mean field approximation

- Independence assumption:

$$\mathcal{Q} = \left\{ q(\theta) : q(\theta) = \prod_{i=1}^{M} q_i(\theta_i), \theta = (\theta_1, ..., \theta_M) \right\}. \quad (12)$$

- Recall $\underline{p}(y; q) = \exp\left(\int q(\theta) \log\left(\frac{p(y,\theta)}{q(\theta)} \right) d\theta \right)$. Then

$$\begin{aligned} & \log \underline{p}(y; q) \\ &= \int q(\theta) \log\left(\frac{p(y, \theta)}{q(\theta)} \right) d\theta \\ &= \int \prod_{i=1}^{M} q_i(\theta_i) \left\{ \log p(y, \theta) - \sum_{i=1}^{M} \log q_i(\theta_i) \right\} d\theta_1 \cdots d\theta_M \quad (13) \end{aligned}$$

Mean field approximation

- continued ...

$$= \int q_1(\theta_1) \left\{ \int \left[\prod_{i \neq 1} q_i(\theta_i) \log p(y, \theta) \right] d\theta_2 \cdots d\theta_M \right\} d\theta_1$$

$$- \int q_1(\theta_1) \log q_1(\theta_1) \, d\theta_1 \underbrace{\int \left[\prod_{i \neq 1} q_i(\theta_i) \right] d\theta_2 \cdots d\theta_M}_{=1} + A_{-1}$$

$$= \int q_1(\theta_1) \left\{ \int \left[\left(\prod_{i \neq 1} q_i(\theta_i) \right) \log p(y, \theta) \right] d\theta_2 \cdots d\theta_M \right\} d\theta_1$$

$$- \int q_1(\theta_1) \log q_1(\theta_1) \, d\theta_1 + A_{-1};$$

$$A_{-1} = - \underbrace{\int q_1(\theta_1) \, d\theta_1}_{=1} \int \left(\sum_{i \neq 1} \log q_i(\theta_i) \right) \left(\prod_{i \neq 1} q_i(\theta_i) \right) d\theta_2 \cdots d\theta_M$$

Mean field approximation: Summary

If the density $q(\theta) = \prod_{i=1}^{M} q_i(\theta_i)$ with $\theta = (\theta_1, ..., \theta_M)$, then

$$\begin{aligned}
\log \underline{p}(y; q) &= \int q(\theta) \log \left(\frac{p(y, \theta)}{q(\theta)} \right) d\theta \\
&= \int q_1(\theta_1) \left\{ \int \left[\left(\prod\nolimits_{i \neq 1} q_i(\theta_i) \right) \log p(y, \theta) \right] d\theta_2 \cdots d\theta_M \right\} d\theta_1 \\
&\quad - \int q_1(\theta_1) \log q_1(\theta_1) \, d\theta_1 + A_{-1},
\end{aligned} \tag{14}$$

where A_{-1} does not involve $q_1(\theta_1)$ and

$$A_{-1} = - \underbrace{\int q_1(\theta_1) \, d\theta_1}_{=1} \int \left(\sum_{i \neq 1} \log q_i(\theta_i) \right) \left(\prod\nolimits_{i \neq 1} q_i(\theta_i) \right) d\theta_2 \cdots d\theta_M$$

Cyclic optimization

- In (14), let

$$\tilde{p}\left(y;\theta_1\right) \propto \exp\left(\int q_2\left(\theta_2\right)\cdots q_M\left(\theta_M\right)\log p\left(y,\theta\right)d\theta_2\cdots d\theta_M\right)$$
$$= \exp\left(E_{-\theta_1}\left[\log p\left(y,\theta\right)\right]\right)$$

- Then $\int \tilde{p}\left(y;\theta_1\right)d\theta_1 dy = 1$ and

$$\log \underline{p}\left(y;q\right) = \int q\left(\theta_1\right)\log\left(\frac{\tilde{p}\left(y;\theta_1\right)}{q_1\left(\theta_1\right)}\right)d\theta_1 + A_{-1}. \tag{15}$$

- Maximizing $\log \underline{p}\left(y;q\right)$ wrt q_1 (of θ_1) by fixing $q_2,\cdots,q_M$ gives

$$\hat{q}_1\left(\theta_1\right) = \tilde{p}\left(\theta_1|y\right) = \frac{\tilde{p}\left(y;\theta_1\right)}{\int \tilde{p}\left(y;\theta_1\right)d\theta_1} \propto \exp\left(E_{-\theta_1}\left[\log p\left(y,\theta\right)\right]\right) \tag{16}$$

due to Jensen's inequality

Cyclic optimization: Summary

- Initialize: $\hat{q}_2(\theta_2), \cdots, \hat{q}_M(\theta_M)$
- Cycle:

$$\hat{q}_1(\theta_1) \leftarrow \frac{\exp\left(E_{-\theta_1}\left[\log p(y,\theta)\right]\right)}{\int \exp\left(E_{-\theta_1}\left[\log p(y,\theta)\right]\right) d\theta_1}$$

$$\vdots$$

$$\hat{q}_M(\theta_M) \leftarrow \frac{\exp\left(E_{-\theta_M}\left[\log p(y,\theta)\right]\right)}{\int \exp\left(E_{-\theta_M}\left[\log p(y,\theta)\right]\right) d\theta_M}$$

until the increase in $\underline{p}(y;q)$ in (13) is negligible, where $E_{-\theta_i}$ denotes expectation wrt $\prod_{j\neq i} q_j(\theta_j)$.

Some shortcomings

- Recall assumption

$$\mathcal{Q} = \left\{ q(\theta) : q(\theta) = \prod_{i=1}^{M} q_i(\theta_i), \theta = (\theta_1, ..., \theta_M) \right\}$$

and optimal solution

$$\hat{q}_i(\theta_i) \propto \exp\left(E_{-\theta_i} [\log p(y, \theta)] \right)$$

- This causes two key issues:
  - Assuming independence among sub-vectors usually leads to inaccuracy in approximating posterior covariance matrix of θ. In general, variational inference underestimates variance of posterior density (partially alleviations for certain models; see Giordano et al. (2018))
  - To obtain $\hat{q}_i(\theta_i)$, need to compute $E_{-\theta_i} [\log p(y, \theta)]$; use "conjugate priors" $\pi(\theta)$ for θ wrt $p(y|\theta)$, so that $E_{-\theta_i} [\log p(y, \theta)]$ has an explicit, analytic expression

VI for frequentists

Variational inference for some frequentists models

- Latent vector u and likelihood

$$l(\theta) = \log p(y;\theta) = \log \int p(y|u;\theta)\, p(u;\theta)\, d\theta \tag{17}$$

  and $l(\theta)$ has no closed form expression
- "Maximum likelihood estimate (MLE)" of θ is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, l(\theta)$$

- Apply variational inference

VI for frequentists

- Pick density $q(u)$ for u. Then

$$p(y,u;\theta) = p(u|y;\theta)\, p(y;\theta), \text{ i.e., } p(y;\theta) = \frac{p(y,u;\theta)}{p(u|y;\theta)}$$

- Further,

$$\begin{aligned} l(\theta) &= \log p(y;\theta) = \log \frac{p(y,u;\theta)}{p(u|y;\theta)} = \log \frac{p(y,u;\theta)}{p(u|y;\theta)} \left(\int q(u)\, du \right) \\ &= \int q(u) \log \frac{p(y,u;\theta)}{p(u|y;\theta)} du \qquad (18) \\ &= \int q(u) \log \frac{p(y,u;\theta)}{q(u)} du + \int q(u) \log \frac{q(u)}{p(u|y;\theta)} du \\ &\geq \underline{l}(\theta;q) = \int q(u) \log \frac{p(y,u;\theta)}{q(u)} du \qquad (19) \end{aligned}$$

VI for frequentists

- Maximizing $\underline{l}(\theta; q)$ is minimizing KL divergence

$$KL(q, p) = \int q(u) \log \frac{q(u)}{p(u|y;\theta)} du \tag{20}$$

- Pick q from $\mathcal{O} = \{q(u;\xi) : \xi \in \Xi\}$. Then

$$\begin{cases} \underline{l}(\theta, \xi; q) = \int q(u;\xi) \log \frac{p(y,u;\theta)}{q(u;\xi)} du \\ KL(q, p) = \int q(u;\xi) \log \frac{q(u;\xi)}{p(u|y;\theta)} du \end{cases}$$

- Just solve $\left(\underline{\hat{\theta}}, \underline{\hat{\xi}}\right) = \text{argmax}_{\theta,\xi}\, \underline{l}(\theta, \xi; q)$ and use $\underline{l}\left(\hat{\theta}, \xi; q\right)$ (see example in Ormerod and Wand (2010))

Frechet regression

Background and motivation

Regression modelling where the response can be

- Directions (modelled by vectors on the unit sphere)
- Information on the directions of neural activity (modelled by positive definite matrices)
- Representation of digital image (modelled by landmark based shape spaces)
- Survival distributions (modelled by densities)
- Wing structures of flies (modelled by graphs)
- Interactions between genes (modelled by graphs)

The Euclidean space

The Euclidean space is so special:

- It is a vector space
- It is a topological space
- It is a normed space

Its vector space structure, topological structure and metric structure are compatible

Examples of a non-Euclidean space

- The unit sphere is not a vector space, even though it inherits topology and norm from its ambient Euclidean space
- The set of positive definite matrices is not a vector space, even though it inherits topology and norm from its ambient Euclidean space
- The set of densities is not a not vector space, and it (usually) does not inherit topology and norm from a Euclidean space
- The set of graphs is not a vector space, and are always be metrizable

Brief review on regression model in Euclidean space

Assume $(X, Y) \sim F$ and $\mu = E(X)$ and $\Sigma = Var(X)$. When both X and $Y \in \mathbb{R}$ are in Euclidean spaces, the regression model is:

$$m(x) = E(Y|X = x) = \beta_0^* + (\beta_1^*)^T (x - \mu)$$

where β_0^* and β_1^* are the solutions

$$(\beta_0^*, \beta_1^*) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \int \left[E(Y|X) - \left\{ \beta_0 + \beta_1^T (x - \mu) \right\} \right]^2 dF_X(x)$$

where

$$E[Y|X] = \int y dF_{Y|X}(x, y)$$

Brief review on regression model in Euclidean space

Derivation of

$$(\beta_0^*, \beta_1^*) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \underbrace{\int \left[E(Y|X) - \left\{ \beta_0 + \beta_1^T (x - \mu) \right\} \right]^2 dF_X(x)}_{\mathcal{L}(\beta_0, \beta_1)}$$

- Part I:

$$\begin{aligned}
\frac{d\mathcal{L}(\beta_0, \beta_1)}{d\beta_0} &= 2 \int \left[E(Y|X) - \left\{ \beta_0 + \beta_1^T (x - \mu) \right\} \right] dF_X(x) \\
&= 2 \int E(Y|x) \, dF_X(x) - 2 \int \left\{ \beta_0 + \beta_1^T (x - \mu) \right\} dF_X(x) \\
&= 2E(Y) - 2\beta_0 - 2\beta_1^T \int (x - \mu) \, dF_X(x)
\end{aligned}$$

So,

$$\frac{d\mathcal{L}(\beta_0, \beta_1)}{d\beta_0} = 0 \Leftrightarrow \beta_0 = E(Y)$$

Brief review on regression model in Euclidean space

Derivation of

$$(\beta_0^*, \beta_1^*) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \underbrace{\int \left[E(Y|x) - \left\{ \beta_0 + \beta_1^T (x - \mu) \right\} \right]^2 dF_X(x)}_{\mathcal{L}(\beta_0, \beta_1)}$$

- Part II:

$$\begin{aligned} \frac{d\mathcal{L}(\beta_0, \beta_1)}{d\beta_1} &= 2 \int (x - \mu) \left[E(Y|x) - \left\{ \beta_0 + (x - \mu)^T \beta_1 \right\} \right] dF_X(x) \\ &= 2 \int (x - \mu) E(Y|x) \, dF_X(x) \\ &\quad - 2 \int (x - \mu) \left\{ \beta_0 + (x - \mu)^T \beta_1 \right\} dF_X \end{aligned}$$

Brief review on regression model in Euclidean space

- Part II:

$$2\int E\left(Y|x\right)\left(x-\mu\right)dF_X\left(x\right)=2\int\left\{\int ydF_{Y|X}\left(x,y\right)\right\}\left(x-\mu\right)dF_X\left(x\right)$$
$$=2E\left(Y\left(X-\mu\right)\right):=2\sigma_{XY}$$

and

$$2\int\left(x-\mu\right)\left\{\beta_0+\left(x-\mu\right)^T\beta_1^T\right\}dF_X\left(x\right)$$
$$=2\int\left(x-\mu\right)\beta_0dF_X\left(x\right)+2\int\left(x-\mu\right)\left(x-\mu\right)^T\beta_1dF_X\left(x\right)$$
$$=2\Sigma\beta_1$$

So

$$\frac{d\mathcal{L}\left(\beta_0,\beta_1\right)}{d\beta_1}=0\Longleftrightarrow\beta_1=\Sigma^{-1}\sigma_{XY}$$

Brief review on regression model in Euclidean space

Derivation of

$$(\beta_0^*, \beta_1^*) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \underbrace{\int \left[E(Y|X) - \left\{ \beta_0 + \beta_1^T (x - \mu) \right\} \right]^2 dF_X(x)}_{\mathcal{L}(\beta_0, \beta_1)}$$

- Solution

$$\beta_0^* = E(Y), \beta_1^* = \Sigma^{-1} \sigma_{XY} = \Sigma^{-1} E(Y(X - \mu))$$

Linear regression in Euclidean space

To derive the above solution, we have utilized $\sigma_{XY} = E(Y(X - \mu))$. However, σ_{XY} cannot be defined this way when

- The space for X does not allow for translation (think about $X - \mu$)
- The space for X does not allow for scalar multiplication (think about $Y(X - \mu)$)
- The space for Y is incompatible with multiplication (think about $Y(X - \mu)$)

How to avoid the need for a vector space structure on Y but still defined linear regression?

Linear regression in non-Euclidean space

To achieve this when the space Ω for Y is only a normed space (but not necessarily a vector space), we need to reply on vector space structure for X. The strategy is to decouple vector space structure from metric structure and interpret the optimization problem differently.

Global Frechet regression

Linear regression in non-Euclidean space

Recall the solution

$$m(x) = E(Y|X = x) = \beta_0^* + (\beta_1^*)^T (x - \mu)$$

with

$$\beta_0^* = E(Y), \beta_1^* = \Sigma^{-1}\sigma_{XY} = \Sigma^{-1}E(Y(X - \mu))$$

Therefore,

$$\begin{aligned} m(x) &= E(Y) + \sigma_{XY}^T \Sigma^{-1}(x - \mu) \\ &= \int y \underbrace{\left\{1 + (z - \mu)^T \Sigma^{-1}(x - \mu)\right\}}_{s(z,x)} dF(z, y) \end{aligned}$$

since

$$\int y(z - \mu)^T \Sigma^{-1}(x - \mu) dF(z, y) = \sigma_{XY}^T \Sigma^{-1}(x - \mu)$$

Linear regression in non-Euclidean space

Recall

$$s(z,x) = 1 + (z-\mu)^T \Sigma^{-1} (x-\mu)$$

Then

$$\begin{aligned}\int s(z,x)\, dF(z,y) &= 1 + \int (z-\mu)^T \Sigma^{-1} (x-\mu)\, dF(z,y) \\ &= 1 + \left(\int (z-\mu)^T dF(z,y) \right) \Sigma^{-1} (x-\mu) \\ &= 1 + 0 \times \Sigma^{-1} (x-\mu)\end{aligned}$$

Therefore,

$$m(x) = E(Y|X=x) = \beta_0^* + (\beta_1^*)^T (x-\mu)$$

is the solution

$$m(x) = \underset{y \in \mathbb{R}}{\operatorname{argmin}}\, E\left[s(X,x)\, d_E^2(Y,y) \right]$$

where d_E is the metric on $\mathbb{R}$.

Frechet regression in Euclidean space

Verification: expansion

$$
\begin{aligned}
& E\left[s\left(X,x\right)d_E^2\left(Y,y\right)\right] \\
& = E\left[d_E^2\left(Y,y\right)\right] - 2yE\left[Y\right] + y^2 + \\
& + E\left[\left(X-\mu\right)^T Y^2\right]\Sigma^{-1}\left(x-\mu\right) - 2y\underbrace{E\left[\left(X-\mu\right)^T Y\right]}_{\sigma_{XY}^T}\Sigma^{-1}\left(x-\mu\right) \\
& + y^2\underbrace{E\left[\left(X-\mu\right)^T\right]\Sigma^{-1}\left(x-\mu\right)}_{=0}
\end{aligned}
$$

where

$$
s\left(z,x\right) = 1 + \left(z-\mu\right)^T\Sigma^{-1}\left(x-\mu\right)
$$

Frechet regression in Euclidean space

Verification: partial derivative

$$\frac{d}{dy} E\left[s\left(X,x\right) d_E^2\left(Y,y\right)\right] = -2E\left(Y\right) + 2y - 2\sigma_{XY}^T \Sigma^{-1}\left(x-\mu\right)$$

so, $E\left[s\left(X,x\right) d_E^2\left(Y,y\right)\right]$ is minimized at

$$y^* = E\left(Y\right) + \sigma_{XY}^T \Sigma^{-1}\left(x-\mu\right)$$

and

$$m\left(x\right) = E\left(Y\right) + \sigma_{XY}^T \Sigma^{-1}\left(x-\mu\right) = \beta_0^* + \left(\beta_1^*\right)^T \left(x-\mu\right)$$

Frechet regression with a Euclidean component

Summary:

$$E\left[s\left(X,x\right)d_E^2\left(Y,y\right)\right]$$
$$=\int \underbrace{s\left(X,x\right)}_{\text{vector space}}\ \underbrace{d^2\left(Y,y\right)}_{\text{metric space}}\ \underbrace{dF\left(X,Y\right)}_{\text{measure space}}$$

The metric space and vector space structures must be compatible to allow the existence of the measure, i.e.,

$$\underbrace{\overbrace{\mathcal{X}}^{\text{vector space}}\otimes\overbrace{\mathcal{Y}}^{\text{metric space}}}_{\text{measure space}}$$

Frechet regression in partial vector space

Setting: $(X, Y) \sim F$ and $\mu = E(X)$ and $\Sigma = Var(X)$, where $X \in \mathbb{R}^p$ and $Y \in (\Omega, d)$ and d is the metric on Ω. Frechet regression is defined as

$$m_{\oplus}(x) = \underset{\omega \in \Omega}{\operatorname{argmin}} M(\omega, x) \quad \text{where} \quad M(\cdot, x) = E\left[s(X, x)\, d^2(Y, \cdot)\right]$$

- Frechet mean:

$$\omega_{\oplus} = \underset{\omega \in \Omega}{\operatorname{argmin}} E\left[d^2(Y, \omega)\right]$$

- Frechet variance:

$$V_{\oplus} = E\left(d^2(Y, \omega_{\oplus})\right)$$

- Frechet regression:

$$m_{\oplus}(x) = \underset{\omega \in \Omega}{\operatorname{argmin}} M_{\oplus}(\omega, x) \quad \text{where} \quad M_{\oplus}(\cdot, x) = E\left[d^2(Y, \cdot) \,|X = x\right]$$

Local Frechet regression

Local Frechet regression

Consider $X \in \mathbb{R}$ and $Y \in \mathbb{R}$. the local linear estimate of $m(x)$ is $\hat{l}(x) = \hat{\beta}_0$, where

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} K_h\left(X_i - x\right) \left(Y_i - \beta_0 - \beta_1^T \left(X_i - x\right)\right)$$

and

$$K_h\left(X_i - x\right) = \frac{1}{h} K\left(\frac{X_i - x}{h}\right)$$

Equivalently

$$\begin{aligned} &\left(\beta_0^*, \beta_1^*\right) \\ &= \underset{\beta_0, \beta_1}{\operatorname{argmin}} \int K_h\left(z - x\right) \left[\int y dF_{Y|X}\left(z, y\right) - \left(\beta_0 + \beta_1^T \left(z - x\right)\right)\right]^2 dF_X\left(z\right) \end{aligned}$$

Local Frechet regression

Let $\mu_j = E\left[K_h\left(X-x\right)\left(X-x\right)^j\right]$ and $r_j = E\left[K_h\left(X-x\right)\left(X-x\right)^j Y\right]$ and $\sigma_0^2 = \mu_0\mu_2 - \mu_1^2$. Claim:

$$\beta_0^* = \frac{\mu_2 r_0 - \mu_1 r_1}{\sigma_0^2}, \beta_1^* = \frac{\mu_0 r_1 - \mu_1 r_0}{\sigma_0^2}$$

Thus $\hat{l}\left(x\right) = \hat{\beta}_0$ can be viewed as an estimate of

$$\begin{aligned}\hat{l}\left(x\right) = \beta_0^* &= \frac{1}{\sigma_0^2}\int yK_h\left(z-x\right)\left[\mu_2 - \mu_1\left(z-x\right)\right]dF\left(z,y\right)\\ &= E\left[s\left(X,x,h\right)Y\right]\end{aligned}$$

where

$$s\left(z,x,h\right) = \frac{1}{\sigma_0^2}K_h\left(z-x\right)\left[\mu_2 - \mu_1\left(z-x\right)\right]$$

Local Frechet regression

Recall $\hat{l}(x) = \hat{\beta}_0$ as an estimate of

$$\hat{l}(x) = \beta_0^* = E[s(X, x, h)\, Y]$$

Since $\int s(z, x, h)\, dF(z, y) = 1$, the estimate $\hat{l}(x)$ is a localized Frechet mean

$$\hat{l}(x) = \underset{y \in \mathbb{R}}{\operatorname{argmin}} E\left[s(X, x, h)\,(Y - y)^2\right]$$

Local Frechet regression:

$$\tilde{l}_{\oplus}(x) = \underset{\omega \in \Omega}{\operatorname{argmin}} E\left[s(X, x, h)\, d^2(Y, \omega)\right]$$

Local Frechet regression

Verification:

$$(\beta_0^*, \beta_1^*)$$

$$= \underset{\beta_0,\beta_1}{\operatorname{argmin}} \underbrace{\int K_h(z-x)\left[\int y dF_{Y|X}(z,y) - \left(\beta_0 + \beta_1^T(z-x)\right)\right]^2 dF_X(z)}_{\mathcal{L}(\beta_0,\beta_1)}$$

with

$$\beta_0^* = \frac{\mu_2 r_0 - \mu_1 r_1}{\sigma_0^2}, \beta_1^* = \frac{\mu_0 r_1 - \mu_1 r_0}{\sigma_0^2}$$

Part I: $g(z) = \int y dF_{Y|X}(z,y)$; expansion

$$\mathcal{L}(\beta_0,\beta_1) = \int K_h(z-x)\left\{g^2(z) - 2g(z)\beta_0 + \beta_0^2\right\} dF_X(z)$$

$$+ \int K_h(z-x)\beta_1(z-x)\left\{-2g(z) + 2\beta_0\right\} dF_X(z)$$

$$+ \int K_h(z-x)\beta_1^2(z-x)^2 dF_X(z)$$

Local Frechet regression

Verification:

$$\mathcal{L}\left(\beta_0, \beta_1\right) = \int K_h\left(z - x\right) \left[\int y dF_{Y|X}\left(z, y\right) - \left(\beta_0 + \beta_1^T\left(z - x\right)\right)\right]^2 dF_X\left(z\right)$$

Part II:

$$\frac{d\mathcal{L}\left(\beta_0, \beta_1\right)}{d\beta_0} = 2\beta_0 \int \underbrace{K_h\left(z - x\right) dF_X\left(z\right)}_{\mu_0} - 2 \underbrace{\int g\left(z\right) K_h\left(z - x\right) dF_X\left(z\right)}_{r_0} + 2\beta_1 \underbrace{\int K_h\left(z - x\right)\left(z - x\right) dF_X\left(z\right)}_{\mu_1}$$

Local Frechet regression

Verification:

$$\mathcal{L}\left(\beta_0, \beta_1\right) = \int K_h\left(z - x\right)\left[\int y dF_{Y|X}\left(z, y\right) - \left(\beta_0 + \beta_1^T\left(z - x\right)\right)\right]^2 dF_X\left(z\right)$$

Part II:

$$\frac{d\mathcal{L}\left(\beta_0, \beta_1\right)}{d\beta_1} = 2\beta_0 \underbrace{\int K_h\left(z - x\right)\left(z - x\right) dF_X\left(z\right)}_{\mu_1}$$

$$- 2 \underbrace{\int K_h\left(z - x\right)\left(z - x\right) g\left(z\right) dF_X\left(z\right)}_{r_1}$$

$$+ 2\beta_1 \underbrace{\int K_h\left(z - x\right)\left(z - x\right)^2 dF_X\left(z\right)}_{\mu_2}$$

Local Frechet regression

$$(\beta_0^*, \beta_1^*)$$

$$= \underset{\beta_0,\beta_1}{\operatorname{argmin}} \underbrace{\int K_h(z-x) \left[\int y dF_{Y|X}(z,y) - \left(\beta_0 + \beta_1^T (z-x) \right) \right]^2 dF_X(z)}_{\mathcal{L}(\beta_0,\beta_1)}$$

with

$$\beta_0^* = \frac{\mu_2 r_0 - \mu_1 r_1}{\sigma_0^2}, \beta_1^* = \frac{\mu_0 r_1 - \mu_1 r_0}{\sigma_0^2}$$

So,

$$\begin{cases} \frac{d\mathcal{L}(\beta_0,\beta_1)}{d\beta_0} = 2\beta_0\mu_0 + 2\beta_1\mu_1 - 2r_0 \\ \frac{d\mathcal{L}(\beta_0,\beta_1)}{d\beta_1} = 2\beta_0\mu_1 + 2\beta_1\mu_2 - 2r_1 \end{cases}$$

By Cramer's rule, the claim holds.

References

Blei, D. M., Ng, A. and Jordan, M. (2003). Latent dirichlet allocation, *J. Mach. Learn. Res.* **3**: 993–1022.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm, *J. R. Statist. Soc. Ser. B* **39**(1): 1–38.

Giordano, R., Broderick, T. and Jordan, M. I. (2018). Covariances, robustness, and variational bayes, *J. Mach. Learn. Res.* **19**(51): 1–49.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations, *Am. Stat.* **64**(2): 140–153.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician, *Ann. Statist.* **12**(4): 1151–1172.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso, *J. R. Statist. Soc. Ser. B* **58**(1): 267–288.