Week 6-1: Sparse PCA; intro to Bayesian analysis

Last time

- PCA, PPCA, Factor analysis

Today

- Sparse PCA
- Introduction to Bayesian analysis

Reference

- Albert, I., S. Donnet, C. Guihenneuc-Jouyaux, ... (2012). Combining Expert Opinions in Prior Elicitation. *Bayesian Analysis* 7:503--532
- Bradley Efron (2013). A 250-year argument: Belief, Behavior, and the bootstrap. *Bulletin (New Series) of the American Mathematical Society*. 50:129--146
- Y. Guan and J. Dy (2009). Sparse Probabilistic Principal Component Analysis. *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, PMLR 5:185-192
- Ning (2021). Spike and slab Bayesian sparse principal component analysis. *arXiv: 2102.00305*
- Robert, C. P. (1994). The Bayesian Choice. 2nd Edition. *Springer Text in Statistics*.
- Ročková, V. and E. I. George (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *JASA* 111:1608–1622.
- F. Yao, H.-G. Müller, and J.-L. Wang (2005). Functional Data Analysis for Sparse Longitudinal Data. *JASA* 100:577--590
- H. Zou, T. Hastie, and Robert Tibshirani (2006). Sparse Principal Component Analysis. *JCGS* 15:265--286

Sparse PCA

Idea: imposing sparsity on loadings. Assume there are s non-zero coordinates in each eigenvector and the remaining $p-s$ coordinates are all 0. Typically, we need $ks\log p \ll n$ (one might can improve it a bit by $ks\log(ep/s)$.

1. Direct sparse approximations

Let $Z_i = U_i D_{ii}$, for each i, we solve

$$\hat\beta = arg\min_{\beta} \|Z_i - X\beta\|^2 + \mathrm{Pen}_\lambda(\beta),$$

then obtain $V_i = \hat\beta/\|\hat\beta\|$.

1. “Self-contained” regression-type criterion (Zou, Hastie, Tibshirani (2006))

Solve

$$(\hat A, \hat B) = arg\min_{A,B} \sum_{i=1}^n \|x_i' - x_i' BA'\|^2 + \mathrm{Pen}_\lambda(\beta), \qquad (1)$$
$$\text{s.t.} \quad A'A = I_{k\times k} \qquad (2)$$

Algorithm:

- Given A, solve B as in the regression setting
- Given B, we minimize $\sum_{i=1}^n \|x_i' - x_i' BA'\|^2$ given $A'A = I_r$. That is, we compute SVD $(X'X)B = UDV'$ and let $\hat A = UV'$.

Other methods:

- Joint-row sparsity for sparse PCA
- Sparse Probabilistic Principal Component Analysis by Guan and Dy (2009).
- Bayseian methods for sparse PCA and factor analysis [Rockova and George (2016); Ning (2021)]
- Functional PCA (Yao, Müller, and Wang (2005))

Sparse PCA in action

SparsePCA is avilable in scikit-learn see here

The method is based on Zou et al (2006)'s paper, the elastic net penalty is added.

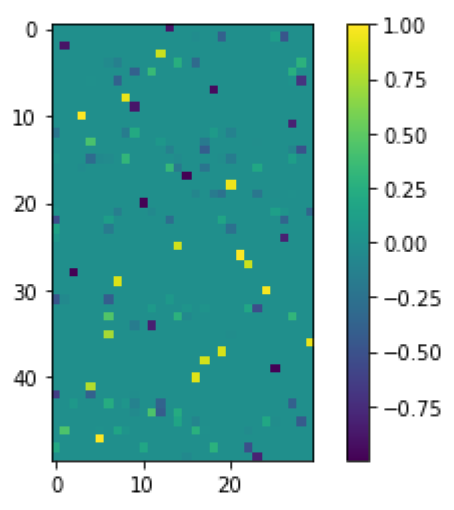
In [55]:
```
from sklearn.decomposition import SparsePCA
spca = SparsePCA(n_components = 30, ridge_alpha = 0.01)
```

In [56]:
```
n = 200
d = 50
mean = np.zeros(d)
cov = np.identity(d)
x = np.random.multivariate_normal(mean, cov, n)

spca.fit(x)
t_spca = spca.transform(x)
p_spca = spca.components_.T
```

In [57]:
```
import matplotlib.pyplot as plt

plt.imshow(p_spca)
plt.colorbar()
plt.show()
```



Bayesian analysis

Bayes rule:

$$\pi(\theta|X) = \frac{P(X|\theta)\pi(\theta)}{P(X)},$$

where $P(X) = \int P(X|\theta)\pi(\theta)d\theta$.

- $P(X|\theta)$: *likelihood*
- $\pi(\theta)$: *prior*
- $\pi(\theta|X)$: *posterior*

Yesterday's posterior is today's prior:

$$\pi(\theta|x_1) \propto f(x_1|\theta)\pi(\theta)$$
$$\pi(\theta|x_1,x_2) \propto f(x_1,x_2|\theta)\pi(\theta) = f(x_2|x_1,\theta)f(x_1|\theta)\pi(\theta) = f(x_2|x_1,\theta)\pi(\theta|x_1)$$
$$\cdots\cdots$$
$$\pi(\theta|x^n, x_{n+1}) \propto f(x_{n+1}, x^n|\theta)\pi(\theta) = f(x_{n+1}|x^n,\theta)\pi(\theta|x^n)$$

Let's watch a video first video

Bradley Efron (2013)

```
Two contending philosophical parties, the Bayesians and the frequentists, have been vying for supremacy over
the past two-and-a-half centuries.The twentieth century was predominantly frequentist, especially in applicat-
ions, but the twenty-first has seen a strong Bayesian revival ... Unlike most philosophical arguments,
this one has important practical consequences. The two philosophies represent competing visions of how
science progress-es and how mathematical thinking assists in that progress.
```

Priors - parametric world

Specifying a prior is the key for conducting Bayesian analysis. It is also the part that has been challenged by frequentism.

1. Conjugate priors

Conjugate priors are commonly used in Bayesian analysis.

Example:

Suppose the observations $x_1,\dots,x_n$ are i.i.d normal $N(\theta,1)$ with unknown mean θ and a known variance 1. The normal prior for θ $\pi(\theta) = N(\mu,\sigma^2)$ is a conjugate prior.

$$\pi(\theta|x_1,\dots,x_n) = \frac{\prod_{i=1}^n f(x_i|\theta)\pi(\theta)}{\int\prod_{i=1}^n f(x_i|\theta)\pi(\theta)d\theta} = N\Big(\frac{\mu/\sigma^2 + \sum_{i=1}^n x_i}{n+1/\sigma^2}, \frac{1}{n+1/\sigma^2}\Big).$$

Definition (P114 of TBC): A family $\mathcal{F}$ of probability distributions on Θ is said to be conjugate for a likelihood function $f(x|\theta)$ if, for every $\pi\in\mathcal{F}$, the posterior distribution $\pi(\theta|x)\in\mathcal{F}$.

Advantage: Convenient. They often lead to a closed form for the posterior, which can be easily used for computation.

Criticisms: often unrealistic. Not necessarily the most robust prior distributions comparing to noninformative priors

Examples of conjugate priors (from TBC)

See more on Wikipedia

2. Noninformative priors

Often, a real prior is hard to obtain, the strategy is to choose a prior so called the non-informative prior such that it will influences the posterior as little as possible.

`Laplace's prior`

The first noninformative prior proposed by Laplace is to choose a uniform prior for the value of parameter. What about $\theta\in\mathbb{R}$? Often one assigns the prior $\pi(\theta)\propto 1$. This prior is improper which can be viewed as the limit of $\mathrm{Unif}(-N,N)$ as $N\to\infty$.

Example:

Consider the prior for $\pi(\theta)\propto 1$ in the previous example, we obtain the posterior

$$\pi(\theta|x_1,\dots,x_n) = N\Big(\frac{\sum_{i=1}^n x_i}{n}, \frac{1}{n}\Big).$$

Criticism: The prior is not invariance under reparameterization. If we switch from θ to $\eta = g(\theta)$, if $\pi(\theta)=1$, the corresponding prior for $\pi(\eta) = |dg^{-1}(\eta)/d\eta|$, in general, is not constant.

`The Jeffreys prior`

The Jefferys noninformative prior is based on the Fisher information matrix given by

$$I(\theta) = -\mathbb{E}_\theta\Big(\frac{\partial^2\log f(X|\theta)}{\partial\theta^2}\Big)$$

The prior is chosen as $\pi(\theta)\propto\sqrt{I(\theta)}$.

This prior is invariant because $I(\theta) = I(g(\theta))(g'(\theta))^2$.

`Other priors` : Reference priors, Haar prior (see Ch 3 of TBC)

3. 'Elicitation from experts' priors

Choose the prior distribution through combining opinions form experts to obtain a valide subjective priors (e.g., Albert et al, 2012).

A quick comment: Different priors can lead to quite different results, but (hopefully) when sample size increases, many of them will agree, and often they even agree with frequentist methods.

There is a large literature on frequentist analysis of Bayesian posteriors (e.g., van der Vaart (1998) Ch.10 of *Asymptotic Statistics*). The goal is to study the limiting behavior of a Bayesian procedure as $n\to\infty$. Surprisingly, by choosing suitable priors, one can show that a Bayesian estimator is consistent (often converges the same limit as some frequentist estimators (e.g., MLE in parametric regular models) and the size of a credible interval matches with the corresponding confidence interval (due to the so called *Bernstein-von phenomenon*).

Hierarchical Bayes and empirical Bayes

`Bayesian hierarchical modelling`

Suppose $y_1,\dots,y_n$ are the observations and one wishes to estimate θ. But there is another parameter ϕ which is unknown, the Bayesian hierarchical model contains the following stages:

1. Likelihood function: $y_j|\theta,\phi \sim f(y_j|\theta,\phi)$
2. $\theta|\phi\sim\pi(\theta|\phi)$
3. $\phi|\pi(\phi)$

Example: In the previous example, suppose $x_i\sim N(\theta,\sigma^2)$ and both θ and σ^2 are unknown, consider the following prior:

$$\pi(\theta|\sigma^2) = N(\mu,\sigma^2), \quad \pi(\sigma^{-2}) = \mathrm{Gamma}(a,b).$$

`Empirical Bayes`

The Empirical Bayes (EB) approach can be seen as an approximation to a fully Bayesian treatment of hierarchical Bayes.

The first major work on EB is by Robbins (1955) (see here) and advocated by Bradley Efron in 1970s.

The EB approach uses data twice. First, it uses data to estimate the prior (often the hyperparameter ϕ). Next, it construct the prior for θ given by $\pi(\theta|\hat\phi)$.