CPSC 540: Machine Learning

Kernel Density Estimation

Mark Schmidt

University of British Columbia

Winter 2020

Last Time: Expectation Maximization

- EM considers learning with observed data O and hidden data H.
- In this case the "marginal" log-likelihood has a nasty form,

$$\log p(O \mid \Theta) = \log\left(\sum_H p(O, H \mid \Theta)\right).$$

- EM applies when "complete" likelihood, $p(O, H \mid \Theta)$, has a nice form.
- EM iterations take the form of a weighted "complete" NLL,

$$\Theta^{t+1} = \underset{\Theta}{\text{argmax}} \left\{\sum_H \alpha_H \log p(O, H \mid \Theta)\right\},$$

where $\alpha_H = p(H \mid O, \Theta^t)$.

  - For mixture models, has a closed-form solution for common distributions.
  - Guarantees monotonic improvment in objective function.
    - Rate of convergence is at least as fast as gradient descent with fixed step size.

EM for MAP Estimation

- We can also use EM for MAP estimation. With a prior on Θ our objective is:

$$\log p(O \mid \Theta) = \log\left(\sum_H p(O, H \mid \Theta)\right) + \log p(\Theta).$$

- EM iterations take the form of a regularized weighted "complete" NLL,

$$\Theta^{t+1} = \underset{\Theta}{\text{argmax}}\left\{\sum_H \alpha_H \log p(O, H \mid \Theta) + \log p(\Theta)\right\},$$

- Now guarantees monotonic improvement in MAP objective.
- This still has a closed-form solution for "conjugate" priors (defined later).

- For mixture of Gaussians with $-\log p(\Theta_c) = \lambda \text{Tr}(\Theta_c)$ for precision matrices Θ_c:
  - Closed-form solution that satisfies positive-definite constraint (no $\log|\Theta|$ needed).

A Non-Parametric Mixture Model

- The classic parametric mixture model has the form

$$p(x^i) = \sum_{c=1}^{k} p(z^i = c) p(x^i \mid z^i = c).$$

- A natural way to define a non-parametric mixture model is

$$p(x^i) = \sum_{j=1}^{n} p(z^i = j) p(x^i \mid z^i = j),$$

where we have one mixture for every training example i.

- Common example: z^i is uniform and $x^i \mid z^i$ is Gaussian with mean x^j,

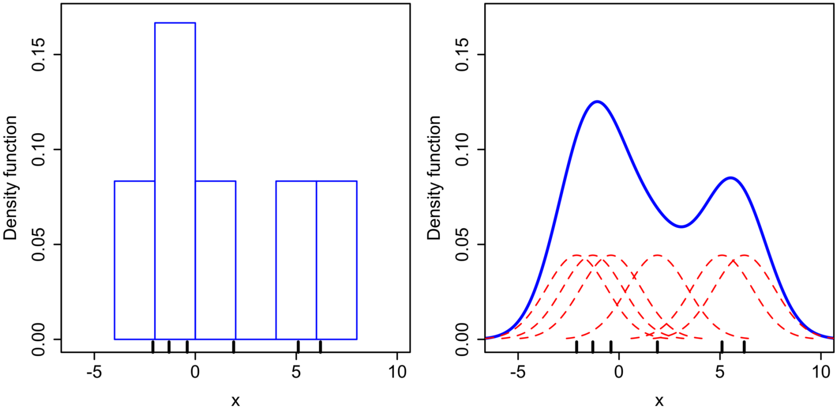
$$p(x^i) = \frac{1}{n} \sum_{j=1}^{n} \mathcal{N}(x^i \mid x^j, \sigma^2 I),$$

and we use a shared covariance $\sigma^2 I$ (σ can be estimated with validation set).

- This is a special case of kernel density estimation (or Parzen window).

Histogram vs. Kernel Density Estimator

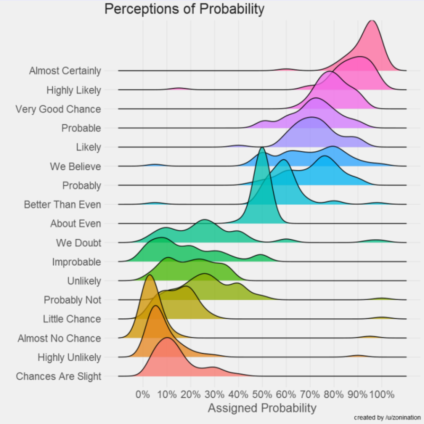
- Think of kernel density estimator as a generalization of a histogram:



https://en.wikipedia.org/wiki/Kernel_density_estimation

Kernel Density Estimator for Visualization

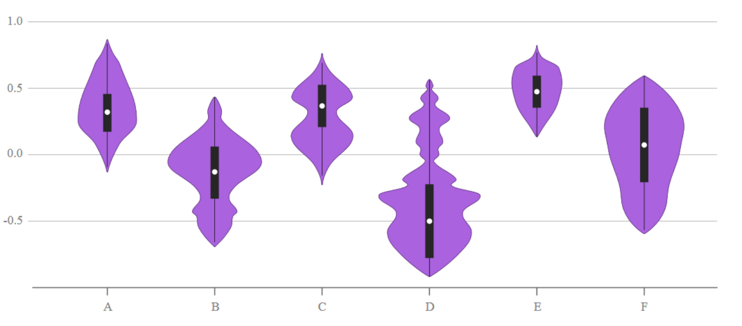
- Visualization of people's opinions about what "likely" and other words mean.



http://blog.revolutionanalytics.com/2017/08/probably-more-probably-than-probable.html

Violin Plot: Added KDE to a Boxplot

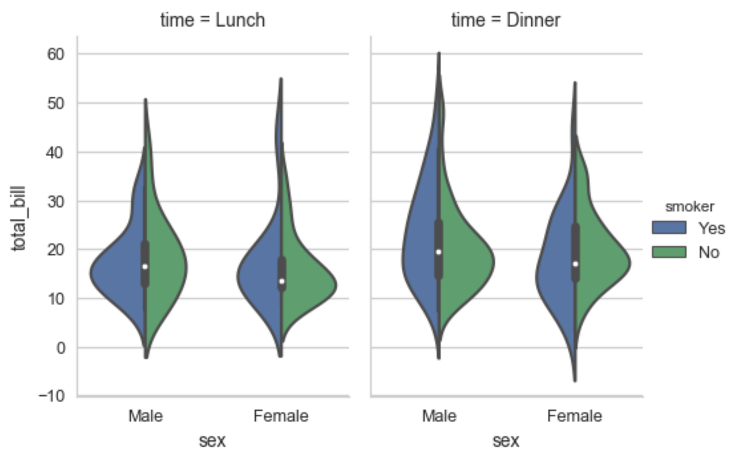
- Violin plot adds KDE to a boxplot:



https://datavizcatalogue.com/methods/violin_plot.html

Violin Plot: Added KDE to a Boxplot

- Violin plot adds KDE to a boxplot:



https://seaborn.pydata.org/generated/seaborn.violinplot.html

Kernel Density Estimation

- The 1D kernel density estimation (KDE) model uses

$$p(x^i) = \frac{1}{n}\sum_{j=1}^{n} k_\sigma \underbrace{(x^i - x^j)}_{r},$$

where the PDF k is the "kernel" and the parameter σ is the "bandwidth".

- In the previous slide we used the (normalized) Gaussian kernel,

$$k_1(r) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{r^2}{2}\right), \quad k_\sigma(r) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{r^2}{2\sigma^2}\right).$$

- Note that we can add a "bandwith" (standard deviation) σ to any PDF k_1, using

$$k_\sigma(r) = \frac{1}{\sigma} k_1\left(\frac{r}{\sigma}\right),$$

from the change of variables formula for probabilities ($|\frac{d}{dr}\left[\frac{r}{\sigma}\right]| = \frac{1}{\sigma}$).

- Under common choices of kernels, KDEs can model any continuous density.

Efficient Kernel Density Estimation

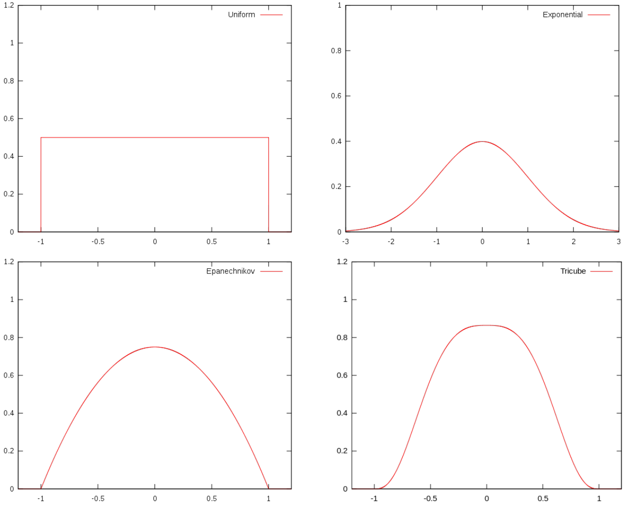
- KDE with the Gaussian kernel is slow at test time:
  - We need to compute distance of test point to every training point.
- A common alternative is the Epanechnikov kernel,

$$k_1(r) = \frac{3}{4}\left(1 - r^2\right)\mathcal{I}\left[|r| \leq 1\right].$$

- This kernel has two nice properties:
  - Epanechnikov showed that it is asymptotically optimal in terms of squared error.
  - It can be much faster to use since it only depends on nearby points (use hashing).
    - You can use hashing to quickly find neighbours in training data.
- It is non-smooth at the boundaries but many smooth approximations exist.
  - Quartic, triweight, tricube, cosine, etc.
- For low-dimensional spaces, we can also use the fast multipole method.

Visualization of Common Kernel Functions

Histogram vs. Gaussian vs. Epanechnikov vs. tricube:



Multivariate Kernel Density Estimation

- The multivariate kernel density estimation (KDE) model uses

$$p(x^i) = \frac{1}{n}\sum_{j=1}^{n} k_A(\underbrace{x^i - x^j}_{r}),$$

- The most common kernel is a product of independent Gaussians,

$$k_I(r) = \frac{1}{(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{\|r\|^2}{2}\right).$$

- We can add a bandwith matrix A to any kernel using

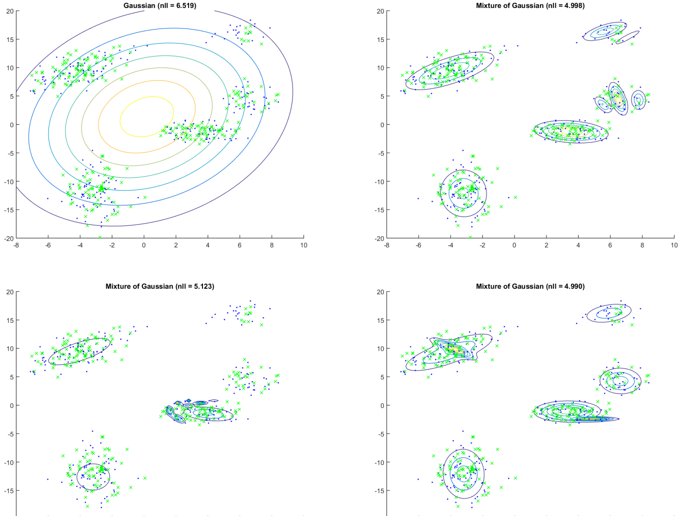
$$k_A(r) = \frac{1}{|A|}k_1(A^{-1}r) \qquad \text{(generalizes } k_\sigma(r) = \frac{1}{\sigma}k_1\left(\frac{r}{\sigma}\right)\text{)},$$

  and in Gaussian case we get a multivariate Gaussian with $\Sigma = AA^T$.

- To reduce number of parameters, we typically:
  - Use a product of independent distributions and use $A = \sigma I$ for some σ.

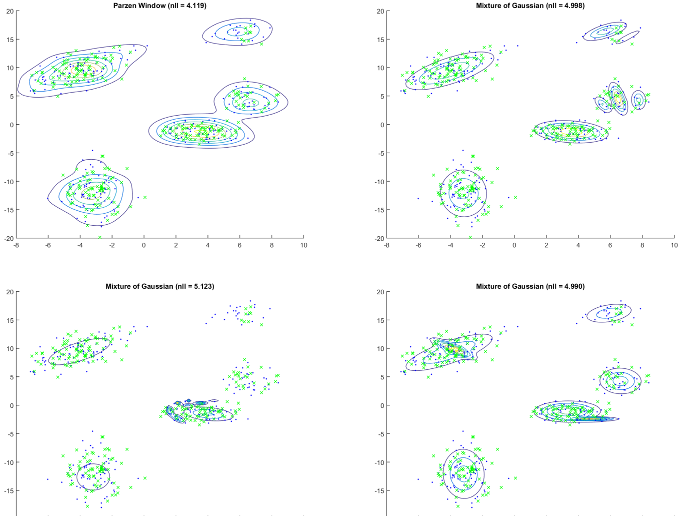
KDE vs. Mixture of Gaussian

- By fixing mean/covariance/k, we don't have to worry about local optima.



KDE vs. Mixture of Gaussian

- By fixing mean/covariance/k, we don't have to worry about local optima.

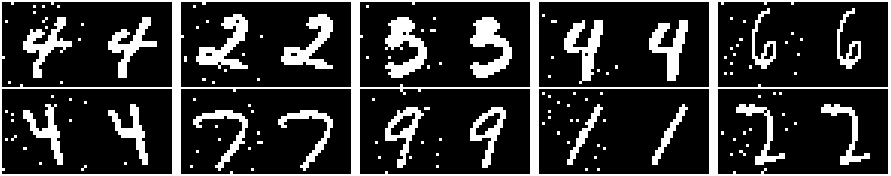


Mean-Shift Clustering

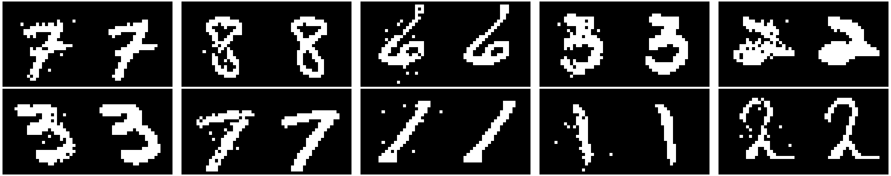
- Mean-shift clustering uses KDE for clustering:
  - Define a KDE on the training examples, and then for test example $\hat{x}$:
    - Run gradient descent to maximize $p(x)$ starting from $\hat{x}$.
  - Clusters are points that reach same local minimum.
- `https://spin.atomicobject.com/2015/05/26/mean-shift-clustering`
- Not sensitive to initialization, no need to choose k, can find non-convex clusters.
- Similar to density-based clustering from 340.
  - But doesn't require uniform density within cluster.
  - And can be used for vector quantization.
- "The 5 Clustering Algorithms Data Scientists Need to Know":
  - `https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68`

Kernel Density Estimation on Digits

- Samples from a KDE model of digits:
  - Sample is on the left, right is the closest image from the training set.



- KDE basically just adds independent noise to the training examples.
  - Usually makes more sense for continuous data that is densely packed.
- A variation with a location-specific variance:



Outline

Continuous Mixture Models

- We've been discussing mixture models where z^i is discrete,

$$p(x^i) = \sum_{z^i=1}^{k} p(z^i)p(x^i \mid z^i = c).$$

- We can also consider mixtures models where z^i is continuous,

$$p(x^i) = \int_{z^i} p(z^i)p(x^i \mid z^i = c)dz^i.$$

- Unfortunately, computing the integral might be hard.
  - But if both probabilities are Gaussian then it's straightforward.

Probabilistic PCA

- In 340 we discussed PCA, which approximates (centered) x^i by

$$x^i \approx W^T z^i.$$

- In probabilistic PCA we assume that

$$x^i \sim \mathcal{N}(W^T z^i, \sigma^2 I), \quad z^i \sim \mathcal{N}(0, I).$$

- We then treat z^i as nuisance parameters,

$$p(x^i \mid W) = \int_{z^i} p(x^i, z^i \mid W) dz^i.$$

- Looks ugly, but this is the marginal of a Gaussian so it's Gaussian.

Probabilistic PCA

- The continuous mixture representation of probabilistic PCA:

$$p(x^i \mid W) = \int_{z^i} p(x^i, z^i \mid W) dz^i = \int_{z^i} p(z^i \mid W) p(x^i \mid z^i, W) dz^i.$$

- After a lot of tedious Gaussian identities and matrix formulas we get (bonus)

$$x^i \mid W \sim \mathcal{N}(0, W^T W + \sigma^2 I),$$

- Regular PCA is obtained as the limit of σ^2 going to 0 (bonus).

- PCA can be viewed as fitting a multivariate Gaussian with a restricted form for Σ.

Generalizations of Probabilistic PCA

- Why do we need a probabilistic interpretation of PCA?
  - Good excuse to play with Gaussian identities and matrix formulas?
- We now understand that PCA fits a Gaussian with restricted covariance:
  - Hope is that $W^TW + \sigma I$ is a good approximation of full covariance X^TX.
  - We can do fancy things like mixtures of PCA models.

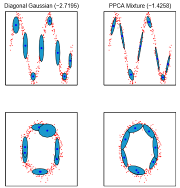


Figure 8: Comparison of an 8-component diagonal variance Gaussian mixture model with a mixture of PPCA model. The upper two plots give a view perpendicular to the major

http://www.miketipping.com/papers/met-mppca.pdf

  - We could consider different $x^i \mid z^i$ distribution (but integrals are ugly).
    - E.g., Laplace or student if you want it to be robust.
    - E.g., logistic or softmax if you have discrete x^i_j.
- Lets us understand connection between PCA and factor analysis.

Factor Analysis

- Factor analysis (FA) is a method for discovering latent-factors.
  - A standard tool and widely-used across science and engineering.
- Historical applications are measures of intelligence and personality traits.
  - Some controversy, like trying to find factors of intelligence due to race.

(without normalizing for relevant factors)

Trait	Description
Openness	Being curious, original, intellectual, creative, and open to new ideas.
Conscientiousness	Being organized, systematic, punctual, achievement-oriented, and dependable.
Extraversion	Being outgoing, talkative, sociable, and enjoying social situations.
Agreeableness	Being affable, tolerant, sensitive, trusting, kind, and warm.
Neuroticism	Being anxious, irritable, temperamental, and moody.

https://new.edu/resources/big-5-personality-traits

- "Big Five" aspects of personality (vs. non-evidence-based Myers-Briggs):
  - https://fivethirtyeight.com/features/most-personality-quizzes-are-junk-science-i-found-one-that-isnt

Factor Analysis

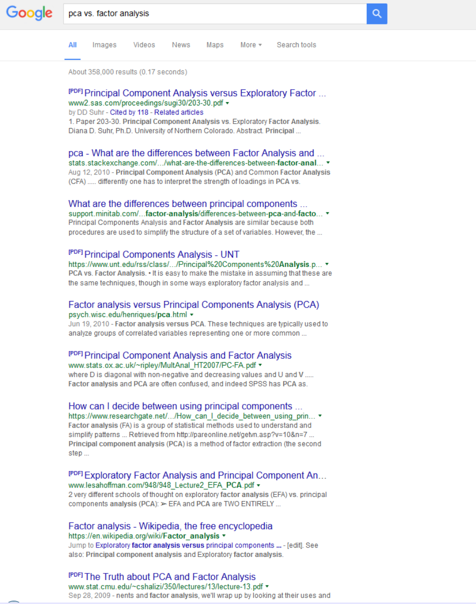
- FA approximates (centered) x^i by

$$x^i \approx W^T z^i,$$

  and assumes z^i and $x^i \mid z^i$ are Gaussian.

- Which should sound familiar...

- Are PCA and FA the same?
  - Both are more than 100 years old.
  - There are many online discussions about whether they are the same.
    - Some software packages run PCA when you call their FA method.
    - Some online discussions claiming they are completely different.


Google
pca vs. factor analysis
All
Images
Videos
News
Maps
More
Search tools
About 358,000 results (0.17 seconds)
[PDF] Principal Component Analysis versus Exploratory Factor ...
www2.sas.com/proceedings/sugi30/203-30.pdf
by DD Suhr - Cited by 118 - Related articles
1. Paper 203-30. Principal Component Analysis vs. Exploratory Factor Analysis. Diana D. Suhr, Ph.D. University of Northern Colorado. Abstract. Principal ...
pca - What are the differences between Factor Analysis and ...
stats.stackexchange.com/.../what-are-the-differences-between-factor-anal...
Aug 12, 2010 - Principal Component Analysis (PCA) and Common Factor Analysis (CFA) ..... differently one has to interpret the strength of loadings in PCA vs.
What are the differences between principal components ...
support.minitab.com/...factor-analysis/differences-between-pca-and-facto...
Principal Components Analysis and Factor Analysis are similar because both procedures are used to simplify the structure of a set of variables. However, the ...
[PDF] Principal Components Analysis - UNT
https://www.unt.edu/rss/class/.../Principal%20Components%20Analysis.p...
PCA vs. Factor Analysis. • It is easy to make the mistake in assuming that these are the same techniques, though in some ways exploratory factor analysis and ...
Factor analysis versus Principal Components Analysis (PCA)
psych.wisc.edu/henriques/pca.html
Jun 19, 2010 - Factor analysis versus PCA. These techniques are typically used to analyze groups of correlated variables representing one or more common ...
[PDF] Principal Component Analysis and Factor Analysis
www.stats.ox.ac.uk/~ripley/MultAnal_HT2007/PC-FA.pdf
where D is diagonal with non-negative and decreasing values and U and V ..... Factor analysis and PCA are often confused, and indeed SPSS has PCA as.
How can I decide between using principal components ...
https://www.researchgate.net/.../How_can_I_decide_between_using_prin...
Factor analysis (FA) is a group of statistical methods used to understand and simplify patterns ... Retrieved from http://pareonline.net/getvn.asp?v=10&n=7 ... Principal component analysis (PCA) is a method of factor extraction (the second step ...
[PDF] Exploratory Factor Analysis and Principal Component An...
www.lesahoffman.com/948/948_Lecture2_EFA_PCA.pdf
2 very different schools of thought on exploratory factor analysis (EFA) vs. principal components analysis (PCA): ➢ EFA and PCA are TWO ENTIRELY ...
Factor analysis - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Factor_analysis
Jump to Exploratory factor analysis versus principal components ... - [edit]. See also: Principal component analysis and Exploratory factor analysis.
[PDF] The Truth about PCA and Factor Analysis
www.stat.cmu.edu/~cshalizi/350/lectures/13/lecture-13.pdf
Sep 28, 2009 - nents and factor analysis, we'll wrap up by looking at their uses and

PCA vs. Factor Analysis

- In probabilistic PCA we assume

$$x^i \mid z^i \sim \mathcal{N}(W^T z^i, \sigma^2 I), \quad z^i \sim \mathcal{N}(0, I),$$

  and we obtain PCA as $\sigma \to 0$.

- In FA we assume

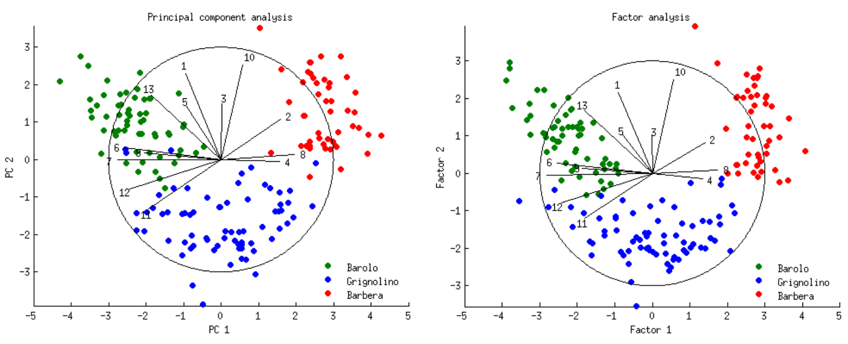
$$x^i \mid z^i \sim \mathcal{N}(W^T z^i, D), \quad z^i \sim \mathcal{N}(0, I),$$

  where D is a diagonal matrix.

- The difference is that you can have a noise variance for each dimension.
  - So FA has extra degrees of freedom in variance of original variables.

PCA vs. Factor Analysis

In practice they usually give pretty similar results:



http://stats.stackexchange.com/questions/1576/what-are-the-differences-between-factor-analysis-and-principal-component-analysi

Remember in 340 that difference with PCA and ISOMAP/t-SNE was huge.

Summary: PCA vs. Factor Analysis

- In probabilistic PCA we assume

$$x^i \mid z^i \sim \mathcal{N}(W^T z^i, \sigma^2 I), \quad z^i \sim \mathcal{N}(0, I),$$

and we obtain PCA as $\sigma \to 0$.

- And factor analysis replaces $\sigma^2 I$ with a diagonal D.

- Differences of FA with PCA:
  - FA is Not affected by scaling individual features.
    - FA doesn't chase large-noise features that are uncorrelated with other features.
  - But unlike PCA, it's affected by rotation of the data (XQ vs. X).
  - No nice "SVD" approach for FA, you can get different local optima.

Independent Component Analysis (ICA)

- Factor analysis has found an enormous number of applications.
  - People really want to find the "factors" that make up their data.
- But even in ideal settings factor analysis can't uniquely identify the true factors.
  - We can rotate W and obtain the same model.
- Independent component analysis (ICA) is a more recent approach.
  - Around 30 years old instead of > 100.
  - Under certain assumptions, it can identify factors.
  - Canonical applications: blind source separation, identifying causal direction.
- It's the only algorithm we didn't cover in 340 from the list of
  "The 10 Algorithms Machine Learning Engineers Need to Know".
- Previous year's material on probabilistic PCA, factor analysis, and ICA here:
  - `https://www.cs.ubc.ca/~schmidtm/Courses/540-W19/L17.5.pdf`

End of Part: Basic Density Estimation and Mixture Models

- We defined the problem of density estimation
  - Computing probability of new examples $\tilde{x}^i$.
- We discussed basic distributions for 1D-case:
  - Bernoulli, categorical, Gaussian.
- We discussed product of independent distributions:
  - Model each feature individually.
- We discussed multivariate Gaussian:
  - Joint Gaussian model of multiple variables.

End of Part: Basic Density Estimation and Mixture Models

- We discussed mixture models:
  - Write density as a convex combination of densities.
  - Examples include mixture of Gaussians and mixture of Bernoullis.
  - Can model multi-modal densities.

- Commonly-fit using expectation maximization.
  - Generic method for dealing with missing at random data.
  - Can be viewed as a "minimize upper bound" method.

- Kernel density estimation is a non-parametric mixture model.
  - Place on mixture component on each data point.
  - Nice for visualizing low-dimensional densities.

Summary

- Kernel density estimation: Non-parametric density estimation method.
  - Allows smooth variations on histograms.
- Probabilistic PCA:
  - Continuous mixture models based on Gaussian assumptions.
  - Factor analysis extends probabilistic PCA with different noise in each dimension.
    - Very similar but not identical to PCA.
- Next time: the sad truth about rain in Vancouver.

Derivation of Probabilistic PCA

- From the probabilistic PCA assumptions we have (leaving out ^{i} superscripts):

$$p(x \mid z, W) \propto \exp\left(-\frac{(x - W^T z)^T (x - W^T z)}{2\sigma^2}\right), \quad p(z) \propto \exp\left(-\frac{z^T z}{2}\right).$$

- Multiplying and expanding we get

$$\begin{aligned} p(x, z \mid W) &= p(x \mid z, W) p(z \mid W) \\ &= p(x \mid z, W) p(z) && (z \perp W) \\ &\propto \exp\left(-\frac{(x - W^T z)^T (x - W^T z)}{2\sigma^2} - \frac{z^T z}{2}\right) \\ &= \exp\left(-\frac{x^T x - x^T W^T z - z^T W x + z^T W W^T z}{2\sigma^2} + \frac{z^T z}{2}\right) \end{aligned}$$

Derivation of Probabilistic PCA

- So the "complete" likelihood satsifies

$$p(x, z \mid W) \propto \exp\left(-\frac{x^Tx - x^TW^Tz - z^TWx + z^TWW^Tz}{2\sigma^2} + \frac{z^Tz}{2}\right)$$
$$= \exp\left(-\frac{1}{2}\left(x^T\left(\frac{1}{\sigma^2}I\right)x + x^T\left(\frac{1}{\sigma^2}W^T\right)z + z^T\left(\frac{1}{\sigma^2}W\right)x + z^T\left(\frac{1}{\sigma^2}WW^T + I\right)z\right)\right),$$

- We can re-write the exponent as a quadratic form,

$$p(x, z \mid W) \propto \exp\left(-\frac{1}{2}\begin{bmatrix}x^T & z^T\end{bmatrix}\begin{bmatrix}\frac{1}{\sigma^2}I & -\frac{1}{\sigma^2}W^T \\ -\frac{1}{\sigma^2}W & \frac{1}{\sigma^2}WW^T + I\end{bmatrix}\begin{bmatrix}x \\ z\end{bmatrix}\right),$$

- This has the form of a Gaussian distribution,

$$p(v \mid W) \propto \exp\left(-\frac{1}{2}(v-\mu)^T\Sigma^{-1}(v-\mu)\right),$$

with $v = \begin{bmatrix}x \\ z\end{bmatrix}$, $\mu = 0$, and $\Sigma^{-1} = \begin{bmatrix}\frac{1}{\sigma^2}I & -\frac{1}{\sigma^2}W^T \\ -\frac{1}{\sigma^2}W & \frac{1}{\sigma^2}WW^T + I\end{bmatrix}$.

Derivation of Probabilistic PCA

- Remember that if we write multivariate Gaussian in partitioned form,

$$\begin{bmatrix} x \\ z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix}\right),$$

then the marginal distribution $p(x)$ (integrating over z) is given by

$$x \sim \mathcal{N}(\mu_x, \Sigma_{xx}).$$

- For probabilistic PCA we assume $\mu_x = 0$, but we partitioned Σ^{-1} instead of Σ.
- To get Σ we can use a partitioned matrix inversion formula,

$$\Sigma = \begin{bmatrix} \frac{1}{\sigma^2}I & -\frac{1}{\sigma^2}W^T \\ -\frac{1}{\sigma^2}W & \frac{1}{\sigma^2}WW^T + I \end{bmatrix}^{-1} = \begin{bmatrix} W^TW + \sigma^2 I & W^T \\ W & I \end{bmatrix},$$

which gives that solution to integrating over z is

$$x \mid W \sim \mathcal{N}(0, W^TW + \sigma^2 I).$$

PCA vs. Probabilistic PCA

- NLL of observed data has the form
$$-\log p(x \mid W) = \frac{n}{2}\text{Tr}(S\Theta) - \frac{n}{2}\log|\Theta| + \text{const.},$$
where $\Theta = (W^TW + \sigma^2 I)^{-1}$ and S is the sample covariance.
- Not convex, but non-global stationary points are saddle points.
- Equivalence with regular PCA:
  - Consider W^T orthogonal so $WW^T = I$ (usual assumption).
  - Using matrix determinant lemma we have
  $$|W^TW + \sigma^2 I| = |I + \frac{1}{\sigma^2}\underbrace{WW^T}_{I}| \cdot |\sigma^2 I| = \text{const.}$$
  - Using matrix inversion lemma we have
  $$(W^TW + \sigma^2 I)^{-1} = \frac{1}{\sigma^2}I - \frac{1}{\sigma^2(\sigma^2+1)}W^TW,$$
  so minimizing NLL maximizes $\text{Tr}(W^TWS)$ as in "analysis" view of PCA.

PCA vs. Factor Analysis

- In probabilistic PCA we assume

$$x^i \mid z^i \sim \mathcal{N}(W^T z^i, \sigma^2 I), \quad z^i \sim \mathcal{N}(0, I),$$

  and we obtain PCA as $\sigma \to 0$.
- In FA we assume

$$x^i \mid z^i \sim \mathcal{N}(W^T z^i, D), \quad z^i \sim \mathcal{N}(0, I),$$

  where D is a diagonal matrix.
- The difference is that you can have a noise variance for each dimension.
- Repeating the previous exercise we get that

$$x^i \sim \mathcal{N}(0, W^T W + D).$$

- So FA has extra degrees of freedom in variance of individual variables.

PCA vs. Factor Analysis

- We can write non-centered versions of both models:
  - Probabilistic PCA:
$$x^i \mid z^i \sim \mathcal{N}(W^T z^i + \mu, \sigma^2 I), \quad z^i \sim \mathcal{N}(0, I),$$
  - Factor analysis:
$$x^i \mid z^i \sim \mathcal{N}(W^T z^i + \mu, D), \quad z^i \sim \mathcal{N}(0, I),$$
    where D is a diagonal matrix.
- A different perspective is that these models assume
$$x^i = W^T z^i + \epsilon,$$
  where PPCA has $\epsilon \sim \mathcal{N}(\mu, \sigma^2 I)$ and FA has $\epsilon \sim \mathcal{N}(\mu, D)$.

Factor Analysis Discussion

- Similar to PCA, FA is invariant to rotation of W,

$$W^TW = W^T \underbrace{Q^TQ}_{I} W = (WQ)^T(WQ),$$

for orthogonal Q.
  - So as with PCA you can't interpret multiple factors as being unique.

- Differences with PCA:
  - Not affected by scaling individual features.
    - FA doesn't chase large-noise features that are uncorrelated with other features.
  - But unlike PCA, it's affected by rotation of the data.
  - No nice "SVD" approach for FA, you can get different local optima.

Orthogonality and Sequential Fitting

- The PCA and FA solutions are not unique.
- Common heuristic:
  1. Enforce that rows of W have a norm of 1.
  2. Enforce that rows of W are orthogonal.
  3. Fit the rows of W sequentially.
- This leads to a unique solution up to sign changes.
- But there are other ways to resolve non-uniqueness (Murphy's Section 12.1.3):
  - Force W to be lower-triangular.
  - Choose an informative rotation.
  - Use a non-Gaussian prior ("independent component analysis").

Scale Mixture Models

- Another weird mixture model is a scale mixture of Gaussians,

$$p(x^i) = \int_{\sigma^2} p(\sigma^2)\mathcal{N}(x^i \mid \mu, \sigma^2) d\sigma^2.$$

- Common choice for $p(\sigma^2)$ is a gamma distribution (which makes integral work):
  - Many distributions are special cases, like Laplace and student t.
- Leads to EM algorithms for fitting Laplace and student t.