Gatsby Theoretical Neuroscience Lectures: Non-Gaussian statistics and natural images Parts III-IV

Aapo Hyvärinen

Gatsby Unit
University College London

Part III: Estimation of unnormalized models

- Often, in natural image statistics, the probabilistic models are unnormalized
  - Major computational problem
- Here, we consider new methods to tackle this problem
- Later, we see applications on natural image statistics

Unnormalized models: Problem definition

- We want to estimate a parametric model of a multivariate random vector $\mathbf{x} \in \mathbb{R}^n$
- Density function f_{norm} is known only up to a multiplicative constant

$$f_{\text{norm}}(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} p_{\text{un}}(\mathbf{x}; \boldsymbol{\theta})$$

$$Z(\boldsymbol{\theta}) = \int_{\boldsymbol{\xi} \in \mathbb{R}^n} p_{\text{un}}(\boldsymbol{\xi}; \boldsymbol{\theta})\, d\boldsymbol{\xi}$$

- Functional form of p_{un} is known (can be easily computed)
- Partition function Z *cannot be computed* with reasonable computing time (numerical integration)
- Here: How to estimate model while avoiding numerical integration?

Examples of unnormalized models related to ICA

- ICA with overcomplete basis simple by

$$f_{\text{norm}}(\mathbf{x}; \mathbf{W}) = \frac{1}{Z(\mathbf{W})} \exp[\sum_i G(\mathbf{w}_i^T \mathbf{x})] \tag{1}$$

- Estimation of second layer in ISA and topographic ICA

$$f_{\text{norm}}(\mathbf{x}; \mathbf{W}, \mathbf{M}) = \frac{1}{Z(\mathbf{W}, \mathbf{M})} \exp[\sum_i G(\sum_i m_{ij} (\mathbf{w}_j^T \mathbf{x})^2)] \tag{2}$$

- Non-Gaussian Markov Random Fields
- ... many more

Previous solutions

- Monte Carlo methods
  - Consistent estimators
    (convergence to real parameter values when sample size $\rightarrow \infty$)
  - Computation very slow (I think)
- Various approximations, e.g. variational methods
  - Computation often fast
  - Consistency not known, or proven inconsistent
- Pseudo-likelihood and contrastive divergence
  - Presumably consistent
  - Computations slow with continuous-valued variables:
    needs 1-D integration at every step, or sophisticated MCMC methods

Content of this talk

- We have proposed two methods for estimation of unnormalized models
- Both methods avoid numerical integration
- First: Score matching (Hyvärinen, JMLR, 2005)
  - Take derivative of model log-density w.r.t. $\mathbf{x}$, so partition function disappears
  - Fit this derivative to the same derivative of data density
  - Easy to compute due to partial integration trick
  - Closed-form solution for exponential families
- Second: Noise-contrastive estimation (Gutmann and Hyvärinen, JMLR, 2012)
  - Learn to distinguish data from artificially generated noise: Logistic regression learns ratios of pdf's of data and noise
  - For known noise pdf, we have in fact learnt data pdf
  - Consistent even in the unnormalized case

Definition of "score function" (in this talk)

- Define model score function $\mathbb{R}^n \to \mathbb{R}^n$ as

$$\psi(\boldsymbol{\xi}; \boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \log f_{\text{norm}}(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_1} \\ \vdots \\ \frac{\partial \log f_{\text{norm}}(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_n} \end{pmatrix} = \nabla_{\boldsymbol{\xi}} \log f_{\text{norm}}(\boldsymbol{\xi}; \boldsymbol{\theta})$$

  where f_{norm} is normalized model density.
- Similarly, define data score function as

$$\psi_{\mathbf{x}}(\boldsymbol{\xi}) = \nabla_{\boldsymbol{\xi}} \log p_{\mathbf{x}}(\boldsymbol{\xi})$$

  where observed data is assumed to follow $p_{\mathbf{x}}(.)$.
- In conventional terminology: Fisher score with respect to a hypothetical location parameter: $f_{\text{norm}}(\mathbf{x} - \boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta} = \mathbf{0}$.

Score matching: definition of objective function

- Estimate by minimizing a distance between model score function $\psi(.;\boldsymbol{\theta})$ and score function of observed data $\psi_{\mathbf{x}}(.)$:

$$J(\boldsymbol{\theta}) = \frac{1}{2}\int_{\boldsymbol{\xi}\in\mathbb{R}^n} p_{\mathbf{x}}(\boldsymbol{\xi})\|\psi(\boldsymbol{\xi};\boldsymbol{\theta}) - \psi_{\mathbf{x}}(\boldsymbol{\xi})\|^2 d\boldsymbol{\xi} \tag{3}$$

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

- This gives a consistent estimator almost by construction
- $\psi(\boldsymbol{\xi};\boldsymbol{\theta})$ does not depend on $Z(\boldsymbol{\theta})$ because

$$\psi(\boldsymbol{\xi};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\xi}} \log p_{\mathrm{un}}(\boldsymbol{\xi};\boldsymbol{\theta}) - \nabla_{\boldsymbol{\xi}} \log Z(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\xi}} \log p_{\mathrm{un}}(\boldsymbol{\xi};\boldsymbol{\theta}) - 0 \tag{4}$$

- *No need* to compute normalization constant Z, non-normalized pdf p_{un} is enough.
- Computation of J quite simple due to theorem below

A computational trick: central theorem of score matching

- In the objective function we have score function of data distribution $\psi_{\mathbf{x}}(.)$. How to compute it?
- In fact, no need to compute it because

Theorem

Assume some regularity conditions, and smooth densities. Then, the score matching objective function J can be expressed as

$$J(\boldsymbol{\theta}) = \int_{\boldsymbol{\xi} \in \mathbb{R}^n} p_{\mathbf{x}}(\boldsymbol{\xi}) \sum_{i=1}^{n} \left[\partial_i \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) + \frac{1}{2} \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta})^2 \right] d\boldsymbol{\xi} + const. \quad (5)$$

where the constant does not depend on $\boldsymbol{\theta}$, and

$$\psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) = \frac{\partial \log p_{un}(\boldsymbol{\xi}; \boldsymbol{\theta})}{\partial \xi_i}, \; \partial_i \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) = \frac{\partial^2 \log p_{un}(\boldsymbol{\xi}; \boldsymbol{\theta})}{\partial \xi_i^2}$$

Simple explanation of score matching trick

- Consider objective function $J(\boldsymbol{\theta})$:

$$\frac{1}{2}\int p_{\mathbf{x}}(\boldsymbol{\xi})\|\psi(\boldsymbol{\xi};\boldsymbol{\theta})\|^2 d\boldsymbol{\xi} - \int p_{\mathbf{x}}(\boldsymbol{\xi})\psi_{\mathbf{x}}(\boldsymbol{\xi})^T\psi(\boldsymbol{\xi};\boldsymbol{\theta})d\boldsymbol{\xi} + \text{const.}$$

- Constant does not depend on $\boldsymbol{\theta}$. First term easy to compute.
- The trick is to use *partial integration* on second term. In one dimension:

$$\int p_x(x)(\log p_x)'(x)\psi(x;\theta)dx = \int p_x(x)\frac{p_x'(x)}{p_x(x)}\psi(x;\theta)dx$$
$$= \int p_x'(x)\psi(x;\theta)dx = 0 - \int p_x(x)\psi'(x;\theta)dx$$

- This is why score function of data distribution $p_{\mathbf{x}}(x)$ disappears!

Final method of score matching

- Replace integration over data density $p_{\mathbf{x}}(.)$ by sample average
- Given T observations $\mathbf{x}(1), \dots, \mathbf{x}(T)$, minimize

$$\tilde{J}(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} \left[\partial_i \psi_i(\mathbf{x}(t); \boldsymbol{\theta}) + \frac{1}{2} \psi_i(\mathbf{x}(t); \boldsymbol{\theta})^2 \right] \quad (6)$$

  where ψ_i is a partial derivative of non-normalized model log-density $\log p_{\text{un}}$, and $\partial_i \psi_i$ a second partial derivative
- Only needs evaluation of some derivatives of the non-normalized (log)-density p_{un} which are simple to compute (by assumption)
- Thus: a new computationally simple and statistically consistent method for parameter estimation

Closed-form solution in the exponential family

- Assume pdf can be expressed in the form

$$\log p_{\mathrm{un}}(\boldsymbol{\xi};\boldsymbol{\theta}) = \sum_{k=1}^{m} \theta_k F_k(\boldsymbol{\xi}) - \log Z(\boldsymbol{\theta}) \tag{7}$$

- Define matrices of partial derivatives:

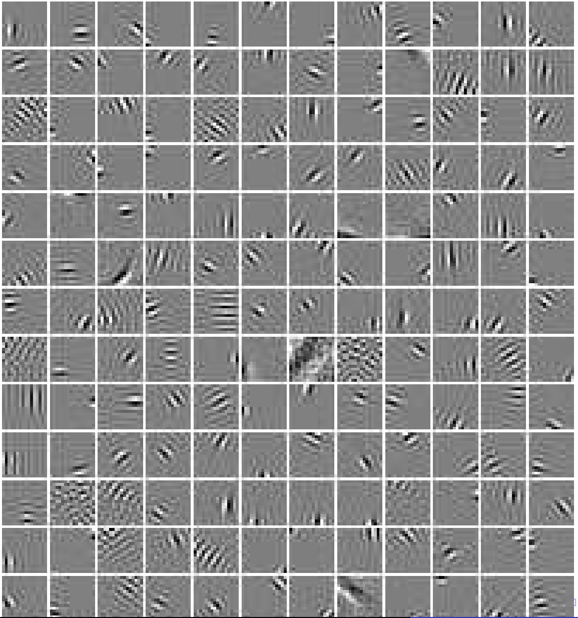
$$K_{ki}(\boldsymbol{\xi}) = \frac{\partial F_k}{\partial \xi_i}, \text{ and } H_{ki}(\boldsymbol{\xi}) = \frac{\partial^2 F_k}{\partial \xi_i^2} \tag{8}$$

- Then, the score matching estimator is given by:

$$\hat{\boldsymbol{\theta}} = -\left[\hat{E}\{\mathbf{K}(\mathbf{x})\mathbf{K}(\mathbf{x})^T\}\right]^{-1} \left(\sum_i \hat{E}\{\mathbf{h}_i(\mathbf{x})\}\right) \tag{9}$$

where $\hat{E}$ denotes the sample average, and the vector $\mathbf{h}_i$ is the i-th column of the matrix $\mathbf{H}$.

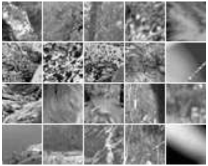
ICA with overcomplete basis



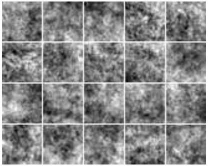
Second method: Noise-contrastive estimation (NCE)

- Train a nonlinear classifier to discriminate observed data from some artificial noise
- To be successful, the classifier must "discover structure" in the data
- For example, compare natural images with Gaussian noise

Natural images



Gaussian noise



Definition of classifier in NCE

- Observed data set $\mathbf{X} = (\mathbf{x}(1), \ldots, \mathbf{x}(T))$ with *un*known pdf $p_\mathbf{x}$
- Generate "noise" $\mathbf{Y} = (\mathbf{y}(1), \ldots, \mathbf{y}(T))$ with known pdf $p_\mathbf{y}$
- Define a nonlinear function (e.g. multilayer perceptron) $g(\mathbf{u}; \theta)$, which models data log-density $\log p_\mathbf{x}(\mathbf{u})$.
- We use logistic regression with the nonlinear function

$$G(\mathbf{u}; \theta) = g(\mathbf{u}; \theta) - \log p_\mathbf{y}(\mathbf{u}). \tag{10}$$

- Well-known developments lead to objective (likelihood)

$$J(\theta) = \sum_t \log \left[h(\mathbf{x}(t); \theta) \right] + \log \left[1 - h(\mathbf{y}(t); \theta) \right]$$

$$\text{where } h(\mathbf{u}; \theta) = \frac{1}{1 + \exp[-G(\mathbf{u}; \theta)]} \tag{11}$$

What does the classifying system do in NCE?

- Theorem:
  - Assume our parametric model $g(\mathbf{u};\theta)$ (e.g. an MLP) can approximate any function.
  - Then, the maximum of classification objective is attained when

$$g(\mathbf{u};\theta) = \log p_{\mathbf{x}}(\mathbf{u}) \tag{12}$$

  where $p_{\mathbf{x}}(\mathbf{u})$ is the pdf of the observed data.
- Corollary: If data generated according to model, i.e. $\log p_{\mathbf{x}}(\mathbf{u}) = g(\mathbf{u};\theta^*)$, we have a *statistically consistent* estimator.
- Supervised learning thus leads to unsupervised estimation of a probabilistic model given by log-density $g(\mathbf{u};\theta)$.

The really important point: NCE estimates unnormalized models

- The maximum of objective function is attained when
  $g(\mathbf{u};\theta) = \log p_{\mathbf{x}}(\mathbf{u})$,
  and there is *no constraint* on g in this optimization problem!
  - In particular, no normalization constraint
    (such as $\int \exp(g(\mathbf{u};\theta))d\mathbf{u} = 1$)
- Even if the family $g(\mathbf{u};\theta)$ is not normalized, the maximum is still attained for the properly normalized pdf
- In practice, normalization constant (partition function) can be estimated like any other parameter
  - For an unnormalized model, add a new parameter c
    $g(\mathbf{u};\theta) \to g(\mathbf{u};\theta) + c$

Choice of noise distribution in NCE

- The noise distribution $p_{\mathbf{y}}$ is an important design parameter.
- We would like to have $p_{\mathbf{y}}$ which fullfills the following:
  1. Easy to sample from
     - But we only need to sample noise once, off-line
  2. Has an analytical expression
     - But we only need to, e.g., normalize it once
  3. It leads to a small mean-squared error of the estimator.
     - This can be analyzed, but optimization not simple
- In practice, we can take Gaussian noise with the same mean and covariance as the data.
- Intuitively, noise should be rather similar to data: classification not too easy

Comparison between score matching and NCE

Computation

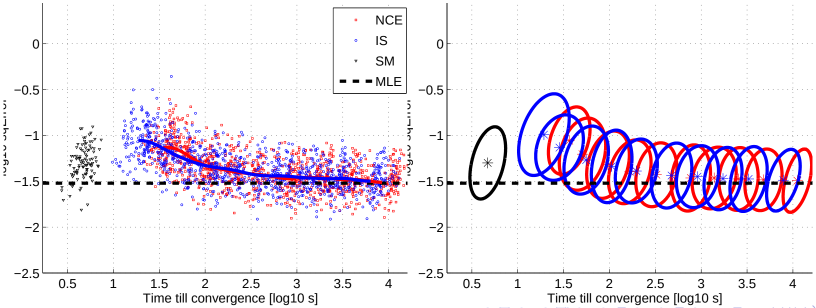
- NCE needs auxiliary noise distribution, while SM does not
- In some models (e.g. multilayer neural networks), SM algebraically difficult
  — Complexity of NCE similar to MLE of normalized model.
- In exponential families, SM particularly simple

Statistics

- Both methods are consistent
- NCE is Fisher-efficient in the limit of infinite noise sample.
- SM probably not Fisher-efficient, but can be shown to have some other optimility properties (Hyvärinen, 2008)
- Noise-contrastive estimation turns out to be closely related to importance sampling (Pihlaja et al, UAI, 2010).
- A general framework can be developed (Gutmann and Hirayama, UAI 2011).

Comparative simulation: computation-statistics trade-off

- Assume potentially infinite data set
- Estimation error limited by computation only
- Compute estimation error vs. computation time for each method
- In NCE, noise sample size determines part of trade-off: For infinite noise sample, Fisher efficient
- Depends strongly on data and model



Conclusion: Estimation of unnormalized models

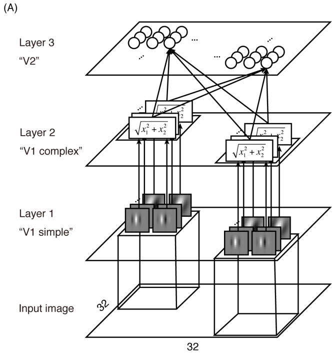
- Unnormalized models important in natural image statistics
- We presented two methods for estimating parameters in unnormalized models
- Unlike typical methods, we avoided numerical integration (or MC methods)
- In score matching, match gradients of log-densities
  —partition function (normalization constant) is completely avoided by taking a derivative
- In noise-contrastive estimation, learn logistic regression to discriminate data from artificial noise
  —partition function can be estimated like any parameter

Part IV: A three-layer model of natural images

- Deep learning is often a black box
- For neurophysiological modelling, we would prefer a network where
  - The role of each unit is clear
  - All cell responses model biological responses
- Instead of blindly stacking many layers on top of each other, we must think about what each layer is doing
- Here: Fix a complex cell model, and estimate another layer by ICA

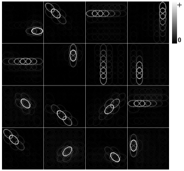
Going towards V2

- Compute *fixed* complex cell outputs for natural images
- Do ICA on complex cell outputs
- A simple model of dependencies in complex cell outputs



Emergence of longer contours

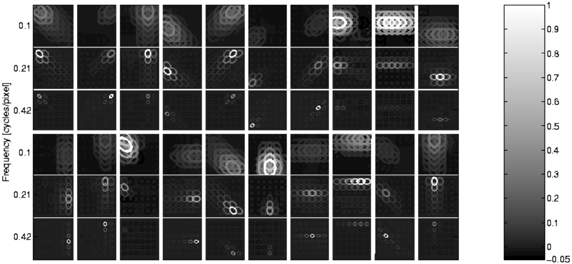
- Hoyer and Hyvärinen (2002) considered a non-negative version of sparse coding
- Main finding: V2 integrates longer contours
- Bayesian inference in the model can model end-stopping etc.



- Cf. "Ultra-long" RF's found by Liu et al (2016).

Emergence of integration over frequencies

- Hyvärinen, Gutmann, and Hoyer (2005) considered several frequency bands (using ordinary ICA)

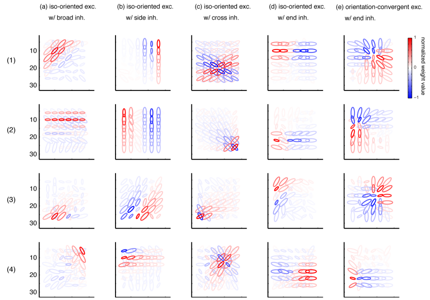


- Each higher-order cell corresponds to 3 frequency displays
- Classic view (of V1) emphasizes *separate* frequency channels
- Integration could be related to sharp edges (Henriksson, Hyvärinen, Vanni, 2009)

Emergence of a variety of RF properties

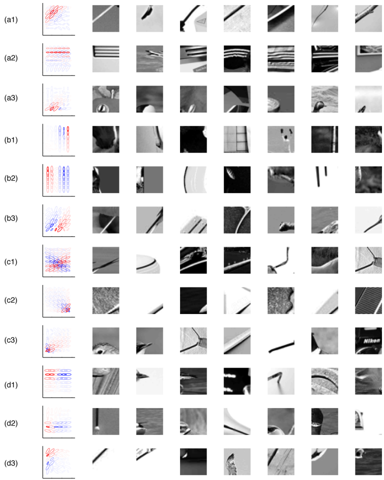
- Hosoya and Hyvärinen (2015) used
  - More densely sampling of orientations
  - Strong PCA dimension reduction
    - One of the simplest possible models of pooling: Works as a simple V1 complex cell model (Hosoya and Hyvärinen, 2016)
  - Overcomplete basis
- Extensive comparison with V2 experiments

Emergence of corner detectors (+ long contours, end-stopping)

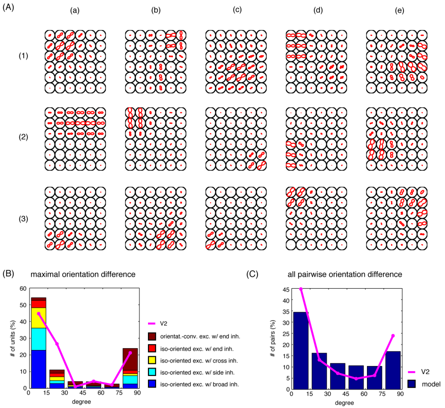


Five principal classes found by Hosoya and Hyvärinen (2015)
Corner detectors (e) are robust, not just a few random gabors

Best natural image patch stimuli



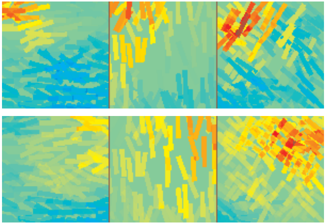
Model reproduces various results on V2

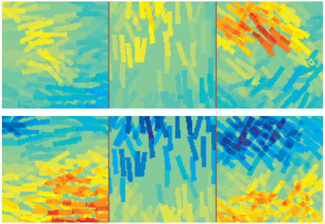


E.g. Spatio-spectral receptive fields similar to Anzai et al (2007)

Can we train all three layers?

- Training all layers (not fixing complex cell model) was done by Gutmann and Hyvärinen (2013)
- Energy-based model trained by noise-contrastive estimation
- Training and interpretation a lot more difficult
- Some receptive fields visualized:





Grand conclusion

- Visual features can be learned from natural images
- Key ingredients in the models
  - Measures of non-gaussian structure:
    - mainly sparsity
  - Non-linearities in processing:
    - invariances as is complex cells by squaring
    - further selectivity in third layer
- We also need suitable methods for estimating the models
  - Maximum likelihood may be computationally infeasible
  - We used score matching and noise-contrastive estimation
- Features often similar to those found in V1,
  or meaningful predictions (third layer)
- Towards predictive theory: New properties emerge (?)