Estimating Densities with Non-Parametric Exponential Families

Lin Yuan
yuanl@purdue.edu
School of ECE
Purdue University

Sergey Kirshner
skirshne@purdue.edu
Dept. of Statistics
Purdue University

Robert Givan
givan@purdue.edu
School of ECE
Purdue University

Abstract

We propose a novel approach for density estimation with exponential families for the case when the true density may not fall within the chosen family. Our approach augments the sufficient statistics with features designed to accumulate probability mass in the neighborhood of the observed points, resulting in a non-parametric model similar to kernel density estimators. We show that under mild conditions, the resulting model uses only the sufficient statistics if the density is within the chosen exponential family, and asymptotically, it approximates densities outside of the chosen exponential family.

1 Non-parametric Exponential Family

Suppose $\boldsymbol{X}$ is a vector of random variables with support $\mathcal{X} \subseteq \mathbb{R}^m$. A distribution for $\boldsymbol{X}$ belongs to the exponential family of distributions with sufficient statistics $\boldsymbol{t} : \mathcal{X} \to \mathcal{H} \subseteq \mathbb{R}^d$, if its probability density has a functional form:[1] $f^E(\boldsymbol{x}|\boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{\lambda})} q(\boldsymbol{x}) \exp\langle \boldsymbol{\lambda}, \boldsymbol{t}(\boldsymbol{x}) \rangle$, where $Z(\boldsymbol{\lambda}) = \int_{\mathcal{X}} q(\boldsymbol{x}) \exp\langle \boldsymbol{\lambda}, \boldsymbol{t}(\boldsymbol{x}) \rangle \, \mathrm{d}\boldsymbol{x} < \infty$ is a partition function, $\boldsymbol{\lambda}$ is a vector of canonical parameters, $q : \mathcal{X} \to \mathbb{R}$ is a base measure, and $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. Assuming q is fixed, let $\mathcal{EF}_{\boldsymbol{t}}$ denote the set of all possible distributions of the form f^E with the set of sufficient statistics $\boldsymbol{t}$.

Given samples $\boldsymbol{x}^{1:n} \triangleq \left(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n\right) \overset{i.i.d}{\sim} f$ where $f : \mathcal{X} \to \mathbb{R}$ is an unknown density with the same support as q. Let $\hat{f}_n : \mathcal{X} \to \mathbb{R}$ be the empirical distribution for $\boldsymbol{x}^{1:n}$, $\hat{f}_n\left(\boldsymbol{x}|\boldsymbol{x}^{1:n}\right) = \frac{1}{n}\sum_{i=1}^n \delta\left(\boldsymbol{x}^i\right)$ where $\delta(\boldsymbol{x})$ is a Dirac delta function. Exponential families can be obtained as a solution to the optimization problem of minimizing the relative entropy subject to matching the moment constraints of the empirical and the estimated distributions.

However, if the true distribution does not fall within the chosen exponential family, $f \notin \mathcal{EF}_{\boldsymbol{t}}$, the estimated model may provide a poor approximation to the true density (e.g. putting most probability mass on the mean when f is multi-modal). Kernel density estimators (KDEs), on the other hand, ensure a small portion of probability mass around the observed data points $\boldsymbol{x}^i$ by a weighted combination of kernel functions: $f_n^{\mathrm{KDE}}\left(\boldsymbol{x}|\boldsymbol{x}^{1:n}\right) = \frac{1}{n}\sum_{i=1}^n K_{\mathbf{H}}\left(\boldsymbol{x}; \boldsymbol{x}^i\right)$, where $K_{\mathbf{H}}\left(\boldsymbol{x}; \boldsymbol{x}^i\right) = |\mathbf{H}|^{-\frac{1}{2}} K\left(\mathbf{H}^{-\frac{1}{2}}\left(\boldsymbol{x} - \boldsymbol{x}^i\right)\right)$. K is a multivariate kernel function, a bounded probability density function on $\mathbb{R}^m$. $K_{\mathbf{H}}$ is a multivariate kernel function with a symmetric positive definite bandwidth matrix $\mathbf{H}$; in this paper, we assume $\mathbf{H} = h^2\mathbf{I}_d$ (assuming $\boldsymbol{x} \in \mathbb{R}^d$). Typically, kernel functions are probability density functions peaked around each data point $\boldsymbol{x}^i$ preserving certain probability mass. For example, the uniform kernel is an indicator function $\mathcal{I}_{\boldsymbol{x}^i}(\boldsymbol{x}) = 1$ if and only if $\left\|\boldsymbol{x} - \boldsymbol{x}^i\right\|_2 < \frac{h}{2}$, and $E_f\left[\mathcal{I}_{\boldsymbol{x}^i}\right]$ is the probability mass inside the $\frac{h}{2}$-ball around $\boldsymbol{x}^i$. Thus,

[1] For notational convenience, we denote $\boldsymbol{X} = \boldsymbol{x}$ by $\boldsymbol{x}$.

KDEs match the mass around each $\boldsymbol{x}^i$ to that of $\hat{f}_n$, therefore are able to approximate any f as $n \to \infty$. Inspired by KDEs, our proposed non-parametric exponential family adds the kernel functions $K_{\mathbf{H}}\left(\boldsymbol{x}; \boldsymbol{x}^i\right) \triangleq t_a^i\left(\boldsymbol{x}\right)$ [2] to the set of sufficient statistics $\boldsymbol{t}\left(\boldsymbol{x}\right)$. We then approximately match $E_f\left[t_a^i\right]$ with $E_{\hat{f}_n}\left[t_a^i\right]$: $\left|E_{\hat{f}_n}\left[t_a^i\right] - E_f\left[t_a^i\right]\right| < \beta_i$. In addition to the canonical parameters $\boldsymbol{\lambda}$ for sufficient statistics, non-parametric exponential families have *augmented* parameters $\boldsymbol{\lambda}_a$ for the augmented statistics $\boldsymbol{t}_a\left(\boldsymbol{x}\right) \triangleq \left[t_a^1 \dots t_a^n\right]$.

The proposed density approximation ($f_n^{NE}\left(\boldsymbol{x}\right)$) is a solution to

$$\begin{aligned} f_n^{NE}\left(\boldsymbol{x}|\boldsymbol{x}^{1:n}\right) &= \underset{f^{NE}\in\mathcal{F}}{\arg\min}\, KL\left(f^{NE} \parallel q\right) \text{ subj to} \\ E_{f_n^{NE}(\boldsymbol{x}|\boldsymbol{x}^{1:n})}\left[\boldsymbol{t}\left(\boldsymbol{x}\right)\right] &= E_{\hat{f}_n(\boldsymbol{x}|\boldsymbol{x}^{1:n})}\left[\boldsymbol{t}\left(\boldsymbol{x}\right)\right], \\ \left|E_{f_n^{NE}}\left[t_a^i\left(\boldsymbol{x}\right)\right] - E_{\hat{f}_n}\left[t_a^i\left(\boldsymbol{x}\right)\right]\right| &\leq \beta_i,\ i = 1, \dots, n. \end{aligned} \tag{1}$$

f_n^{NE} falls within the generalized MaxEnt framework [1]:

$$\begin{aligned} f\left(\boldsymbol{x}\right) &= \frac{1}{Z\left(\boldsymbol{\lambda}, \boldsymbol{\lambda}_a\right)} q\left(\boldsymbol{x}\right) \exp\left[\left\langle\boldsymbol{\lambda}, \boldsymbol{t}\left(\boldsymbol{x}\right)\right\rangle + \left\langle\boldsymbol{\lambda}_a, \boldsymbol{t}_a\left(\boldsymbol{x}\right)\right\rangle\right] \\ Z\left(\boldsymbol{\lambda}, \boldsymbol{\lambda}_a\right) &= \int_{\mathcal{X}} q\left(\boldsymbol{x}\right) \exp\left[\left\langle\boldsymbol{\lambda}, \boldsymbol{t}\left(\boldsymbol{x}\right)\right\rangle + \left\langle\boldsymbol{\lambda}_a, \boldsymbol{t}_a\left(\boldsymbol{x}\right)\right\rangle\right] \mathrm{d}\,\boldsymbol{x}. \end{aligned} \tag{2}$$

Let $\boldsymbol{s}\left(\boldsymbol{x}\right) \triangleq \left(\boldsymbol{t}\left(\boldsymbol{x}\right), \boldsymbol{t}_a\left(\boldsymbol{x}\right)\right)$ and $\boldsymbol{\theta} \triangleq \left(\boldsymbol{\lambda}, \boldsymbol{\lambda}_a\right)$ be a combined set of statistics and parameters, respectively, for the augmented model. A specific set of parameter values for the distribution in (2) satisfying the constraints in (1) can be found by maximizing the penalized log-likelihood

$$l\left(\boldsymbol{\theta}\right) = \left\langle\boldsymbol{\theta}, \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}\left(\boldsymbol{x}^i\right)\right\rangle - \ln Z\left(\boldsymbol{\theta}\right) - \sum_{i=1}^{n} \beta_i \left|\lambda_a^i\right|. \tag{3}$$

1.1 Theoretical Properties

The proofs appear in [2].

Theorem 1.1. *Suppose a vector of random variables $\boldsymbol{X}$ with support on $\mathcal{X}$ has a density $f \in \mathcal{EF}_{\boldsymbol{t}}$ with features $\boldsymbol{t} : \mathcal{X} \to \mathcal{H} \subseteq \mathbb{R}^d$ and a vector of canonical parameters $\boldsymbol{\lambda} \in \mathcal{C} \in \mathbb{R}^d$. Suppose $\boldsymbol{x}^1, \dots, \boldsymbol{x}^n \triangleq \boldsymbol{x}^{1:n}$ is a sequence of i.i.d. random vectors drawn from f. Let $f_n^{NE}\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}^{1:n}\right) \in \mathcal{NEF}_{\boldsymbol{s}}$ be the MLE solution of (1), $\hat{\boldsymbol{\theta}}_n = \left(\tilde{\boldsymbol{\lambda}}_n, \tilde{\boldsymbol{\lambda}}_{a,n}\right)$, with all $\beta_i = \beta > 0$, $i = 1, \dots, n$. Assuming 1. $\mathcal{X}$ is compact, 2. $\boldsymbol{t}$ is continuous, 3. $\mathcal{EF}_{\boldsymbol{t}}$ is a family of uniformly equicontinuous functions w.r.t $\boldsymbol{x}$, 4. Kernel K has bounded variation and has a bandwidth parameter $\mathbf{H}$ such that the series $\sum_{n=1}^{\infty} e^{-\gamma n|\mathbf{H}|}$ converges for every positive value of γ, then as $n \to \infty$, $\tilde{\lambda}_{a,n}^i \xrightarrow{p} 0, \forall i = 1, \dots, n$ and $\tilde{\boldsymbol{\lambda}}_n \xrightarrow{p} \boldsymbol{\lambda}$.*

Theorem 1.1 shows that if the true distribution falls within the exponential family, then as sample size increases, the estimated density from the non-parametric exponential family will have vanishing reliance on the augmented parameters.

Theorem 1.2. *Given a probability density function $f\left(\boldsymbol{x}\right) : \mathcal{X} \to \mathbb{R}$, let $f_n^{NE}\left(\boldsymbol{x}|\hat{\theta}_n, \boldsymbol{x}^{1:n}\right) \in \mathcal{NEF}_{\boldsymbol{s}}$ be a solution satisfying (1). If 1. f is uniformly continuous on $\mathcal{X}$, 2. $K_{\mathbf{H}}\left(\boldsymbol{x}\right)$ is uniformly continuous on $\mathcal{X}$, 3. $\sup_{\boldsymbol{x}\in\mathcal{X}} K_{\mathbf{H}}\left(\boldsymbol{x}\right) < \infty$, 4. $\lim_{\|\boldsymbol{x}\|\to\infty} K_{\mathbf{H}}\left(\boldsymbol{x}\right) \prod_{i=1}^{m} \boldsymbol{x}_i = 0$, 5. $\lim_{n\to\infty} |\mathbf{H}|^{\frac{1}{2}} = 0$, 6. $\lim_{n\to\infty} n\,|\mathbf{H}|^{\frac{1}{2}} = \infty$, then $f_n^{NE}\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}^{1:n}\right) \xrightarrow{p} f\left(\boldsymbol{x}\right)$ pointwise on $\mathcal{X}$.*

Theorem 1.2 indicates the weak consistency of the non-parametric exponential family density estimator. Thus our proposed non-parametric approach can be used to approximate densities which are not from exponential families.

[2] We omit $\mathbf{H}$(h in univariate kernels) from t_a^i for the simplicity of notation. It is a tuning parameter that may be set globally for all $i = 1, \dots, n$.

Algorithm 1 Non-Parametric Exponential Family Coordinate Descent

INPUT: Samples $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n \in \mathbb{R}^d$, sufficient statistics $\boldsymbol{t} : \mathcal{X} \to \mathcal{H}$, augmented features $t_a^i : \mathcal{H} \to \mathbb{R}$, $i = 1, \ldots, n$, ℓ_1 regularization parameters $\boldsymbol{\beta}$
OUTPUT: MLE $\boldsymbol{\theta} = \left(\lambda^1, \ldots, \lambda^d, \lambda_a^1, \ldots, \lambda_a^n\right)$
Initialize $\boldsymbol{\theta}^{(0)}$
Compute the sufficient statistics $E_{\hat{f}_n(\boldsymbol{x}|\boldsymbol{x}^{1:n})}\left[\boldsymbol{t}\left(\boldsymbol{x}\right)\right]$
repeat
  iteration $k = k + 1$, $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k-1)}$
  for $i = 1, \ldots, d$ **do**
    $g_i^{(k)} = E_{\hat{f}_n}\left[t^i\left(\boldsymbol{x}\right)\right] - E_{f_n^{NE}\left(\boldsymbol{x}|\theta^{(k)}\right)}\left[t^i\left(\boldsymbol{x}\right)\right]$
    Perform line search along $g_i^{(k)}$ to update $\lambda^{i,(k)}$
  end for
  for $j = 1 \ldots n$ **do**
    Solve two equations for λ_a^j (denote the solutions $\lambda_a^{j,-}$ and $\lambda_a^{j,+}$, respectively):
    $E_{f_n^{NE}\left(\boldsymbol{x}|\boldsymbol{\theta}^{(k)}\right)}\left[t_a^j\left(\boldsymbol{x}\right)\right] = E_{\hat{f}_n}\left[t_a^j\left(\boldsymbol{x}\right)\right] - \beta_j$ and $E_{f_n^{NE}\left(\boldsymbol{x}|\boldsymbol{\theta}^{(k)}\right)}\left[t_a^j\left(\boldsymbol{x}\right)\right] = E_{\hat{f}_n}\left[t_a^j\left(\boldsymbol{x}\right)\right] + \beta_j$
    set $\lambda_a^{j,(k)} = \lambda_a^{j,-}$ if $\lambda_a^{j,-} > 0$, set $\lambda_a^{j,(k)} = \lambda_a^{j,+}$ if $\lambda_a^{j,+} < 0$, and set $\lambda_a^{j,(k)} = 0$ otherwise
  end for
until convergence
return $\boldsymbol{\theta}^{(k)}$

1.2 Estimating Parameters for Non-Parametric Exponential Families

Recently there have been a number of methods developed for optimization of convex non-smooth functions, some of them specifically aimed at log-linear problems such as (3) [e.g., 3–5]. We employed a coordinate descent algorithm similar to the SUMMET algorithm of [1] (see Algorithm 1), primarily, due to its simplicity. Other possible approaches can be employed as well and may end up more efficient for this formulation.

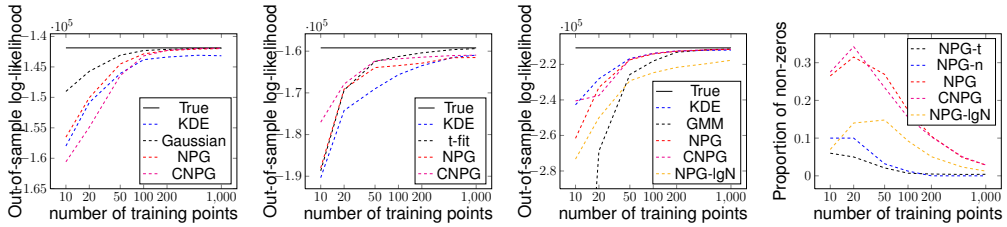
The proposed algorithm iterates between optimizing canonical parameters $\boldsymbol{\lambda}$ (by setting $E_{\hat{f}_n}\left[\boldsymbol{t}\left(\boldsymbol{x}\right)\right] = E_{f_n^{NE}\left(\boldsymbol{x}|\boldsymbol{\theta}^{(k)}\right)}\left[\boldsymbol{t}\left(\boldsymbol{x}\right)\right]$) and sequentially optimizing the augmented parameters $\boldsymbol{\lambda}_a$ so that the Karush-Kuhn-Tucker conditions [e.g., 6] are satisfied:

$$E_{\hat{f}_n}\left[t_a^i\left(\boldsymbol{x}\right)\right] - E_{f_n^{NE}(\boldsymbol{x}|\boldsymbol{\theta})}\left[t_a^i\left(\boldsymbol{x}\right)\right] \in \begin{cases} \{\beta_i\} & \lambda_a^i > 0, \\ \{-\beta_i\} & \lambda_a^i < 0, \\ (-\beta_i, \beta_i) & \beta_a^i = 0. \end{cases}$$

Algorithm 1 belies the inherent difficulty of: (1) calculation of the partial derivative $g_i^{(k)}$, and (2) an implicit search procedure to update $\boldsymbol{\lambda}_a^{j,(k)}$, both involve calculating intractable integrals. If the support is low-dimensional and the mass is contained in a small volume, then the partition function (and thus the gradient) can be computed by numerical integration (quadrature). Alternatively, a common approach to MLE with an intractable partition function $Z\left(\boldsymbol{\theta}\right)$ is Markov Chain Monte Carlo MLE [MCMC-MLE, 7]. For example, the time complexity at each iteration k is $O(Sn^2)$, where S is the number of Monte-Carlo samples we choose to use. However, we believe developments in optimization [e.g. 3, 5] will help us find an efficient solution.

2 Experimental Evaluation

We illustrate the behavior of the proposed non-parametric density estimator matching first and second order moment constraints (NPGaussian, i.e. $\boldsymbol{t}(x) = \left(x, x^2\right)$) in the univariate setting. Normal density (in $\mathcal{EF}$, $\mathcal{N}(0,1)$), t-distribution (not in $\mathcal{EF}$, df=6) and a mixture of two normals (not in $\mathcal{EF}$, $\frac{1}{2}\mathcal{N}(-3,1) + \frac{1}{2}\mathcal{N}(3,1)$) are used for simulating i.i.d samples. We vary the sample size from 10 to 1000 for training and compute the out-of-sample likelihood with an evaluation set of 100000 samples for testing. Adaptive Gauss-Lobatto quadrature is used for numerical integration. We compared the performance of our non-parametric approach, the model from the true functional family,



(a) (b) (c) (d)

Figure 1: Estimating simple one dimensional densities. Results are averaged over 20 runs. The x axis is in log scale. (a) Normal distribution (b) t distribution (c) Mixed normal distribution (d) Number of non-zero $\boldsymbol{\lambda}_a$s. In Figure(a,b,c), `NPG`: NPGaussian with $O(1/\sqrt{n})$ schedule, `NPG-lgN`: NPGaussian with $O(1/\log(n))$ schedule, `CNPG`: constrained NPGaussian with true global moment statistics. In (d), `NPG-t`: number of non-zeros for estimating t distribution with NPG, `NPG-n`: number of non-zeros for estimating normal distribution with NPG, `NPG`: number of non-zeros for estimating mixed normal distribution with NPG, `CNPG`: number of non-zeros for estimating mixed normal distribution with CNPG, `NPG-lgN`: number of non-zeros for estimating mixed normal distribution with NPG-lgN.

and another non-parametric approach (KDE). There are two sets of tuning parameters, bandwidth h and the box constraint parameter β, assumed to be the same for all $i = 1, \ldots, n$. β was set according to a fixed schedule $\boldsymbol{\beta}(n) = O(1/\sqrt{n})$. h (both for KDE and for our approach) was determined based on cross-validated log-likelihood.[3] Gaussian kernel function is used for NPGaussian and for KDE. When estimating normal density, the NPGaussian model (NPG) quickly converges to the normal density with the augmented parameters vanishing as suggested by Theorem 1.1 (Figure 1(a,d)). We also consider the case when the true sufficient statistics are given to us (constrained NPGaussian, CNPG). The CNPG model shows improvement over NPGaussian for small n. However, as the training sample size increases, both CNPG and NPGaussian show similar performance as the moment constraints $\boldsymbol{t}(x)$ are more accurately approximated. NPGaussian gives comparable performance with KDE when estimating densities $\notin \mathcal{EF}_{\boldsymbol{t}}$ (Figure 1(b,c)). We also experimented with $O(1/\log(n))$ regularization schedule (NPG-lgN) for βs when estimating the mixed normal distribution. As n increase, NPG-lgN becomes sparser than NPGaussian (Fig. 1(d)) and gives a worse performance than KDE (Figure 1(c)), which agrees with Theorem 1.2.

References

[1] M. Dudik, S. Phillips, and R. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, 8: 1217–1260, Jun 2007.

[2] L. Yuan, S. Kirshner, and R. Givan. Estimating Densities with Non-Parametric Exponential Families. *ArXiv e-prints*, June 2012. URL `http://arxiv.org/abs/1206.5036`.

[3] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. In S. Sra, S. Nowozin, and S. J. Wright., editors, *Optimization for Machine Learning*, chapter 2. MIT press, 2011.

[4] T. Wu and K. Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, March 2008.

[5] S. Shalev-Shwartz and A. Tewari. Stochastic methods for l_1 regularized loss minimization. *Journal of Machine Learning Research*, 12:1865–1892, June 2011.

[6] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer, 2nd edition, 2006.

[7] C. J. Geyer and E. A. Thompson. Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society*, 54(3):pp. 657–699, 1992.

[3] Interestingly, unlike KDE, cross-validation implies the choice of h does not impact NPGaussian much.