A finite element method for density estimation with Gaussian process priors

Markus Hegland [1]
markus.hegland@anu.edu.au

Australian National University

November 1, 2011

[1] with Michael Griebel, Bonn

Abstract

Probability densities are a major tool in exploratory statistics and stochastic modelling. I will talk about a numerical technique for the estimation of a probability distribution from scattered data using exponential families and a maximum a-posteriori approach with Gaussian process priors. Using Cameron-Martin theory, it can be seen that density estimation leads to a nonlinear variational problem with a functional defined on a reproducing kernel Hilbert space. This functional is strictly convex. A dual problem based on Fenchel duality will also be given. The (original) problem is solved using a Newton-Galerkin method with damping for global convergence. In this talk I will discuss some theoretical results relating to the numerical solution of the variational problem and the results of some computational experiments. A major challenge is of course the curse of dimensionality which appears when high-dimensional probability distributions are estimated.

Outline

Outline

The problem

- data: $D = x_1, \ldots, x_n \in X$ drawn randomly from some unknown probability distribution p
- probability density model: $p(x \mid u) = \exp(u(x) - \gamma(u))$ where $\gamma(u)$ is such that $\int p(x \mid u)\, dx = 1$
- estimation problem: for given data $x_1, \ldots, x_n$ find $\hat{u}(x)$ such that $p(x \mid \hat{u})$ approximates underlying density
- likelihood:

$$p(D \mid u) = \exp\left(\sum_{i=1}^{n} u(x_i) - n\,\gamma(u)\right)$$

  choose $\hat{u}$ such that likelihood large
- if X finite: maximum likelihood method
- problem underdetermined if X infinite

The MAP approach

- prior for u: Gaussian probability measure ν over space of functions = Gaussian process prior
- posterior based on likelihood $p(D \mid u)$:

$$d\mu = p(D \mid \cdot)\, d\nu$$

  is a well defined measure if $p(D \mid \cdot) \in L_1(\nu)$
- maximum a-posteriori (MAP) method:
  estimate $\hat{u}$ is mode of μ

Outline

The Cameron-Martin approach

- for infinite X the measure μ does not have a density
- consider shifted measures $\mu_h(A) = \mu(A + h)$ and

$$\mu_h(A) = \int_A \frac{d\mu_h}{d\mu}\, d\mu, \quad h \in H$$

  where H is the Cameron-Martin space (RKHS)
- can show that for the MAP method

$$\frac{d\mu_h}{d\mu} = p(D \mid \cdot)\, \frac{d\nu_h}{d\nu}$$

- assume that the Radon-Nikodym derivative $\frac{d\mu_h}{d\mu}(u)$ is a continuous function of u

The mode of a probability measure

- a function u_1 is *more likely* than u_2 if
  - $h = u_2 - u_1 \in H$
  - there exists an open set $\mathcal{U}$ containing u_1 such that

$$\mu_h(A) < \mu(A), \quad \text{for all measurable } A \subset \mathcal{U}$$

- this is defines a partial order on the set of functions u and a mode of μ is a maximal point
- characterisation of a mode: if

$$d\mu_h/d\mu(u_1) < 1$$

then by continuity of the Radon-Nikodyn derivative, there exists an open $\mathcal{U}$ with $d\mu_h/d\mu(v) < 1$ for $v \in \mathcal{U}$ and so $\mu_h(A) = \int_A d\mu_h/d\mu(v)\, d\mu < \mu(A)$ for measurable $A \subset \mathcal{U}$

A non-normalised density

- chain rule for Radon-Nikodym derivatives:

$$\frac{d\mu_{h_1+h_2}}{d\mu}(u) = \frac{d\mu_{h_1}}{d\mu}(u+h_2)\,\frac{d\mu_{h_2}}{d\mu}(u), h_1, h_2 \in H$$

  as $\frac{d\mu_{h_1+h_2}}{d\mu_{h_2}}(u) = \frac{d\mu_{h_1}}{d\mu}(u+h_2)$

- replace u, h_1, h_2 by $0, u, h$ and set $\rho(u) = \frac{d\mu_u}{d\mu}(0)$:

$$\frac{d\mu_h}{d\mu}(u) = \frac{\rho(u+h)}{\rho(u)}, \quad u, h \in H$$

- for finite X and density f on has $\rho(u) = f(u)/f(0)$
- in our case

$$\rho(u) = p(D \mid u)\,\exp(-\|u\|^2_{CM})$$

  where $\|\cdot\|_{CM}$ is the Cameron-Martin norm

The variational problem

- u is a mode of μ if $\rho(u) \geq \rho(u+h)$ for $h \in H$
- in MAP $\rho(u) = \exp(-n\,j(u))$ with

$$j(u) = \frac{1}{n}\|u\|_{CM}^2 - \frac{1}{n}\sum_{i=1}^{n} u(x_i) + \gamma(u)$$

where $\gamma(u) = \log \int_X \exp(u(x))\,dx$

- with $\|u\|_H^2 = \frac{2}{n}\|u\|_{CM}^2$ we have

$$j(u) = \frac{1}{2}\|u\|_H^2 + \log \int_X \exp(u(x))\,dx - \frac{1}{n}\sum_{i=1}^{n} u(x_i).$$

- MAP estimator: $\hat{u} = \operatorname{argmin}_{u \in H} j(u)$

Outline

Properties of j

- H is reproducing kernel Hilbert space (RKHS): $|u(x)| \leq C_H \|u\|_H$ and thus

$$\frac{1}{2}\|u\|_H^2 - 2C_H\|u\|_H \leq j(u) \leq \frac{1}{2}\|u\|_H^2 + 2C_H\|u\|_H$$

  so j is bounded on H and thus *proper*
- if $\|u_k\|_H \to \infty$ then $j(u_k) \to \infty$ thus j is *coercive*
- j is lower semicontinuous
- j is stricly convex

Theorem

The functional $j(u)$ has exactly one minimum in H

Characterisation of the minimum of j

- Gateaux derivative

$$\langle \nabla j(u), v \rangle = \lim_{\tau \to 0} \frac{j(u + \tau v) - j(u)}{\tau}$$

- Gateaux derivative of j using Taylor expansion of log and exp

$$\langle \nabla j(u), v \rangle = (u, v)_H + \frac{\int_X e^{u(x)} v(x)\, dx}{\int_X e^{u(x)}\, dx} - \frac{1}{n} \sum_{i=1}^{n} v(x_i)$$

and so $\|\nabla j(u)\| \leq (1 + 2C_H)\|u\|_H$

Theorem

u is a minimum of j if and only if $\langle \nabla j(u), v \rangle = 0$ for all $v \in H$

An approximate kernel density estimator

- if $\int_X dx = 1$ and $\int u(x)\,dx = 0$ then $\log \int_X \exp(u(x))\,dx \approx \frac{\|u\|^2}{2}$ and so

$$j(u) \approx \frac{1}{2}\left(\|u\|_H^2 + \|u\|^2\right) - \frac{1}{n}\sum_{i=1}^{n} u(x_i)$$

- there exists $g_x \in H$ and

$$u(x) = \frac{1}{n} g_{x_i}$$

- density $p(x \mid u) = \exp(u(x) - \gamma(u)) \approx 1 + u(x)$ and so

$$p(x \mid u) \approx \frac{1}{n}\sum_{i=1}^{n}(1 + g_{x_i})$$

Dual functionals

- let $E : H \to L_2(X)$ be continuous embedding and

$$j_0(u) = \frac{1}{2}\|u\|_H^2 - \frac{1}{n}\sum_{i=1}^{n} u(x_i) \quad \text{and} \quad j_1(z) = \log \int_X e^{z(x)}\, dx$$

  for $u \in H$ and $z \in L_2(X)$ and so $j(u) = j_0(u) + j_1(Eu)$
- the duals are for $u^* \in H$ and $z^* \in L_2(X)$ with $z^*(x) > 0$ a.e. and $\int_X z^*(x)\, dx = 1$ ($j_1^*(z^*) = \infty$ else)

$$\begin{aligned} j_0^*(u^*) &= \frac{1}{2}\|u^*\|_H^2 + \frac{1}{n}\sum_{i=1}^{n} u^*(x_i) + \frac{1}{2}\left\|\frac{1}{n}\sum_{i=1}^{n} k_{x_i}\right\|_H^2 \\ j_1^*(z^*) &= \int_X z^*(x)\log z^*(x)\, dx \end{aligned}$$

Fenchel duality

$$\min_{u\in H} j_0(u) + j_1(u) = -\min_{z\in L_2(X)} j_0^*(-Ez^*) + j_1^*(z^*)$$

furthermore, one has the connection between u and z^*:

$$u(x) = \int_X k(x,y)z^*(y)\,dy - \frac{1}{n}\sum_{i=1}^n k(x,x_i)$$

and jointly, (u, z^*) minimise the augmented functional

$$\phi(u,z^*) = \frac{1}{2}\|u\|_H^2 + \int_X z^*(x)\log z^*(x)\,dx$$

minimum characterised as the solution of the integral equation

$$\log z^*(x) + \int k(x,y)z^*(y)\,dy = \frac{1}{n}\sum_{i=1}^n (k(x,x_i) - 1)$$

Outline

A strong formulation

- Galerkin equations $\langle \nabla j(u), v \rangle = 0$ for $v \in H$

$$\langle \nabla j(u), v \rangle = (u, v)_H + \frac{\int_X e^{u(x)} v(x)\, dx}{\int_X e^{u(x)}\, dx} - \frac{1}{n} \sum_{i=1}^{n} v(x_i) = 0$$

- Strong equations $F(u) = 0$ from $(F(u), v)_H = \langle \nabla j(u), v \rangle$

$$F(u) = u + \frac{\int_X e^{u(x)} k_x\, dx}{\int_X e^{u(x)}\, dx} - \frac{1}{n} \sum_{i=1}^{n} k_{x_i}$$

where the integral $w = \int_X e^{u(x)} k_x\, dx$ is weakly defined by

$$(w, v) = \int_X e^{u(x)} v(x)\, dx$$

Fréchet derivative of F and energy norm

- Fréchet derivative DF is linear operator with

$$F(u+v) - F(u) - DF(u)v = o(\|v\|_H)$$

thus

$$DF(u)v = v + \frac{\int_X e^{(u(x)} k_x v(x)\, dx}{\int_X e^{u(x)}\, dx} - \frac{\int_X e^{u(x)v(x)}\, dx \int_X e^{u(x)} k_x\, dx}{(\int_X e^{u(x)}\, dx)^2}$$

- local energy norm $a(v, w \mid u) = (v, DF(u)w)_H$

$$a(v, w \mid u) = (v, w)_H + \frac{\int_X e^u vw\, dx}{\int_X e^u\, dx} - \frac{\int_X e^u v\, dx \int_X e^u w\, dx}{\left(\int_X e^u\, dx\right)^2}$$

A stochastic interpretation of the local energy norm

- recall that $p(x \mid u) = e^{u(x)} \Big/ \int_X e^{u(y)}\, dy$
- expectation and covariance of p:

$$\begin{aligned} E(v \mid u) &= \int_X p(x \mid u) v(x)\, dx \\ \mathrm{Cov}(v, w \mid u) &= \int_X p(x \mid u)(v(x) - E(v \mid u))(w(x) - E(w \mid u))\, d \end{aligned}$$

- local energy

$$a(v, w \mid u) = (v, w)_H + \mathrm{Cov}(v, w \mid u)$$

Newton iterative method

- $u^{k+1} = u^k - DF(u^k)^{-1}F(u^k), \quad k = 1, 2, \ldots$
  convergence by Newton-Kantorovich and Newton-Mysoskikh:
- As $a(v, v; u) \geq \|v\|_H$

$$\|DF(u)^{-1}\| = \sup_{v \in H} \frac{\|v\|_H^2}{a(v, v; u)} \leq 1$$

- by Cauchy-Schwarz, mean-value theorem and quotient rule:

$$\|DF(u_2) - DF(u_1)\| = \sup_{v,w \in H} \frac{a(v, w; u_2) - a(v, w; u_1)}{(v, w)_H}$$

$$\leq \sup_{v \in H} \frac{\int_T |p(t; u_2) - p(t; u_1)| v(t)^2 \, dt}{\|v\|_H^2} \leq 2C_H^3 \|u_2 - u_1\|_H$$

The Newton-Galerkin method

- method established by Deuflhard (2004)

$$u^{k+1} = u^k - DF(u^k)^{-1}(F(u^k) - r^k)$$

with the Galerkin condition

$$(r^k, u^{k+1} - u^k)_H = 0$$

- global convergence achieved by damping

$$u^{k+1} = u^k - \lambda_k \, DF(u^k)^{-1}(F(u^k) - r^k)$$

- solution in (finite dimensional) subspace $V \subset H$

$$(DF(u^k)s^k + F(u^k), v) = 0, \quad v \in V$$

where $s^k \in V$ and $u^{k+1} = u^k + \lambda_k s^k$

Outline

Errors

- error in the probabilities (L_1 or L_2 for $\int_X dx = 1$)

$$\begin{aligned} |P(A) - \hat{P}(A)| &\leq \int_X |p(x \mid u) - p(x \mid \hat{u})| \, dx \\ &\leq \sqrt{\int_X (p(x \mid u) - p(x \mid \hat{u}))^2 \, dx} \end{aligned}$$

- error in the expected values

$$|E(v \mid u) - E(v \mid \hat{u})| \leq \|v\|_H \|p(\cdot \mid u) - p(\cdot \mid \hat{u})\|_{H^*}$$

- how to compute the dual norm

$$\|e\|_{H^*} = \sup_{v \in H} \frac{\int_X e(x)v(x) \, dx}{\|v\|_H} = \int_X k(x,y)e(x)e(y) \, dxdy$$

A reference solution

- given probability $p > 0$ find $u_r \in V$ such that

$$\frac{\int_X e^{u_r(x)} v(x)\,dx}{\int_X e^{u_r(x)}\,dx} = \int_X p(x)v(x)\,dx, \quad v \in V$$

- u_r minimises the functional

$$j_{\text{exact}}(u) = \log \int_0^1 e^{u(x)}\,dx - \int_0^1 p(x)u(x)\,dx$$

"maximum likelihood fit"

The onedimensional case

- $X = [0, 1]$ and $H = H_0^1[0, 1]$ with norm

$$\|u\|_H = \sqrt{\frac{\alpha}{n}\left(\int_0^1 u'(x)^2\,dx + \beta^2 \int_0^1 u(x)^2\,dx\right)}$$

  for some $\beta \geq 0$ and $\alpha > 0$

- reproducing kernel k:

$$k(x, y) = k(y, x) = \frac{\sinh(\beta(1-x))\sinh(\beta y)}{\beta \sinh(\beta)}, \quad y \leq x$$

Recovering the 1D normal distribution

truncated normal distribution with mean 0.5 and variance 0.1
$\alpha = 0.0002$ and $\beta = 1$, piecewise linear approximation, L_1 and L_2 errors

l	$e^{(1)}_{1,l}$	$\frac{e^{(1)}_{1,l}}{e^{(1)}_{1,l+1}}$	$e^{(1)}_{2,l}$	$\frac{e^{(1)}_{2,l}}{e^{(1)}_{2,l+1}}$	$e^{(3)}_{1,l}$	$\frac{e^{(3)}_{1,l}}{e^{(3)}_{1,l+1}}$	$e^{(3)}_{2,l}$	$\frac{e^{(3)}_{2,l}}{e^{(3)}_{2,l+1}}$
2	2.1e-1	–	4.3e-1	–	4.9e-1	–	6.0e-1	–
3	5.1e-2	4.2	1.0e-1	4.4	2.8e-1	1.8	3.9e-1	1.6
4	1.2e-2	4.0	2.5e-2	4.1	1.3e-1	2.2	2.0e-1	2.0
5	3.1e-3	4.0	6.1e-3	4.0	6.4e-2	2.0	1.0e-1	1.9
6	7.8e-4	4.0	1.5e-3	4.0	3.1e-2	2.1	5.1e-2	2.1
7	2.0e-4	4.0	3.8e-4	4.0	1.4e-2	2.1	2.4e-2	2.1
8	5.0e-5	4.0	9.6e-5	4.0	6.2e-3	2.4	1.0e-2	2.4
9	1.2e-5	4.0	2.4e-5	4.0	2.1e-3	3.0	3.5e-3	3.0

left: maximum likelihood projection, $O(h^2)$ error as $u \in C^2[0,1]$
right: our estimator and 1000 data points, $O(h^{3/2-\varepsilon})$ as $\hat{u} \in H^{3/2-\varepsilon}$

Data on grid and Old Faithful data

compared against very fine grid solution in L_1 and L_2

l	$e^{(2)}_{1,l}$	$\frac{e^{(2)}_{1,l}}{e^{(2)}_{1,l+1}}$	$e^{(2)}_{2,l}$	$\frac{e^{(2)}_{2,l}}{e^{(2)}_{2,l+1}}$	$e^{(4)}_{1,l}$	$\frac{e^{(4)}_{1,l}}{e^{(4)}_{1,l+1}}$	$e^{(4)}_{2,l}$	$\frac{e^{(4)}_{2,l}}{e^{(4)}_{2,l+1}}$
2	4.2e-2	–	5.4e-2	–	1.2	–	1.4	–
3	2.3e-2	1.9	3.2e-2	1.7	7.8e-1	1.6	1.1	1.4
4	5.6e-3	4.1	9.0e-3	3.5	4.0e-1	2.0	6.2e-1	1.7
5	9.1e-4	6.1	1.7e-3	5.4	2.0e-1	2.0	3.3e-1	1.9
6	2.9e-4	3.1	4.5e-4	3.7	1.0e-1	2.0	1.9e-1	1.7
7	7.2e-5	4.0	1.1e-4	4.0	5.3e-2	2.0	1.0e-1	1.9
8	1.8e-5	4.0	2.8e-5	4.0	2.2e-2	2.4	4.4e-2	2.3
9	4.1e-6	4.4	6.5e-6	4.3	7.6e-3	2.9	1.5e-2	2.9

left: normal distribution, data truncated to grid points for $l = 6$ grid, $O(h^2)$
right: Old Faithful geyser data, $O(h^{3/2})$

The 2D case

$H = H_0^1 \times H_0^1$ tensor product of 1D spaces with

$$\|u\|_H = \frac{\alpha}{n}\sqrt{\int_0^1 \int_0^1 \left(u_{x,y}^2 + \beta^2 u_x^2 + \beta^2 u_y^2 + \beta^4 u^2\right)\, dx\, dy}$$

l	$e^{(1)}_{1,l}$	$\frac{e^{(1)}_{1,l}}{e^{(1)}_{1,l+1}}$	$e^{(1)}_{2,l}$	$\frac{e^{(1)}_{2,l}}{e^{(1)}_{2,l+1}}$	$e^{(3)}_{1,l}$	$\frac{e^{(3)}_{1,l}}{e^{(3)}_{1,l+1}}$	$e^{(3)}_{2,l}$	$\frac{e^{(3)}_{2,l}}{e^{(3)}_{2,l+1}}$
1	9.13e-01	–	5.04e+00	–	8.08e-02	–	1.33e-01	–
2	3.16e-01	2.89e+00	1.38e+00	3.67e+00	7.75e-02	1.04e+00	1.23e-01	1.08e+00
3	7.08e-02	4.46e+00	2.41e-01	5.70e+00	2.91e-02	2.66e+00	6.15e-02	1.99e+00
4	1.73e-02	4.10e+00	5.98e-02	4.04e+00	7.18e-03	4.06e+00	1.63e-02	3.78e+00
5	4.34e-03	3.98e+00	1.51e-02	3.97e+00	1.47e-03	4.88e+00	3.42e-03	4.76e+00
6	1.05e-03	4.12e+00	3.80e-03	3.96e+00				

maximum likelihood projection and estimator

Outline

A Newton-Galerkin Opticom method

- Newton Galerkin $u_{n+1} = u_n + \Delta u_n, \quad \Delta u_n$ minimises

$$J(\Delta u) = \frac{1}{2} H_{u_n}(\Delta u, \Delta u) + (F(u_n), \Delta u)_H$$

- Sparse grid space $V = \sum_j V^{(j)}$
- Sparse grid combination technique $\Delta u_n = \sum_{j=1}^{k} c_j \Delta u_n^{(i)}$ where components $\Delta u_n^{(j)}$ minimise $J(\Delta u)$ over $V^{(j)}$
- Opticom method: choose combination coefficients c_j to minimise $J(\sum_{j=1}^{k} c_j \Delta u_n^{(j)}) \Rightarrow$ descent method, converges to sparse grid solution, not some combination approximation

Errors of sparse grid approximation for 2D case

approximation of the normal distribution: maximum likelihood projection and our estimator

l	$e^{(1)}_{1,l}$	$\frac{e^{(1)}_{1,l}}{e^{(1)}_{1,l+1}}$	$e^{(1)}_{2,l}$	$\frac{e^{(1)}_{2,l}}{e^{(1)}_{2,l+1}}$	$e^{(3)}_{1,l}$	$\frac{e^{(3)}_{1,l}}{e^{(3)}_{1,l+1}}$	$e^{(3)}_{2,l}$	$\frac{e^{(3)}_{2,l}}{e^{(3)}_{2,l+1}}$
1	1.42e+00	–	–	–	8.05e-02	–	1.33e-01	–
2	3.12e-01	4.55e+00	1.16e+00	–	7.97e-02	1.01e+00	1.27e-01	1.04e+00
3	7.37e-02	4.23e+00	2.44e-01	4.75e+00	3.11e-02	2.56e+00	6.39e-02	1.99e+00
4	1.94e-02	3.81e+00	6.34e-02	3.85e+00	9.63e-03	3.23e+00	1.89e-02	3.38e+00
5	4.92e-03	3.93e+00	1.60e-02	3.96e+00	3.13e-03	3.08e+00	6.14e-03	3.08e+00
6	1.23e-03	4.00e+00	4.17e-03	3.83e+00	8.04e-04	3.89e+00	1.72e-03	3.56e+00

Conclusion

- The application of the MAP method to infinite dimensional parameter settings requires the application of the Cameron–Martin density as the ordinary density does not exist
- Gaussian process priors lead to well-posed problems which can be solved by a Newton-Galerkin method
- For high dimensions we have developed a variant which uses the Opticom method and has the advantages of the combination technique but converges to the sparse grid solution

References


Vladmir I. Bogachev
Gaussian Measures
American Mathematical Society, 1998


Eberhard Zeidler
Nonlinear functional analysis and its applications III
Springer, 1985


Michael Griebel and Markus Hegland
A finite element method for density estimation with Gaussian process priors
SIAM J. Num. Anal., 2010