Density Estimation with Orthogonal Polynomials

Georg Heimel

October 4, 2017

Abstract

This is the abstract.

Chapter 1

Orthogonal Polynomials

1.1 The univariate case

We take *orthogonal* polynomials $\phi_i(x)$ of a single, continuous variable to exhibit the following properties.

1. Their index $i = 0 \ldots \infty$ stands for the highest power of x occurring in the polynomial, also called its *degree*.

2. The polynomials are orthogonal on some interval (a, b) with respect to some measure $w(x)$ in the sense that:

$$\int_a^b \mathrm{d}x\, \phi_i(x)\phi_j(x)w(x) = \delta_{ij} \tag{1.1}$$

3. As such they form a complete basis of all square-integrable $L^2(\mathbb{R})$ functions on the interval (a, b), that is, if $f(x)$ be such a function, then

$$f(x) = \sum_{j=0}^{\infty} c_j\, \phi_j(x)w(x) \tag{1.2}$$

with constant coefficients c_j, which are determined by multiplying eq. 1.2 from the left with $\phi_i(x)$ and integrating over all x.

$$\int_a^b \mathrm{d}x\, \phi_i(x)f(x) = \sum_{j=0}^{\infty} c_j \int_a^b \mathrm{d}x\, \phi_i(x)\phi_j(x)w(x)$$

Using eq. 1.1 on the right-hand side, we then find:

$$c_i = \int_a^b \mathrm{d}x\, \phi_i(x)f(x) \tag{1.3}$$

4. We can approximate $f(x)$ by a *truncated* series

$$f(x) \approx \sum_{j=0}^{J} c_j \, \phi_j(x) w(x) \tag{1.4}$$

with $J < \infty$ and we can make this approximation systematically better by increasing J.

1.2 The bivariate case

Orthogonal polynomials $\Phi_{ij}(x, y)$ of *two* continuous variables, x and y, are simple the tensor product $\phi_i(x) \otimes \phi_j(y)$. As such, they inherit the properties of their univariate components.

1. Their indices, $i, j = 0 \ldots \infty$ stand for the highest powers of x and y occurring in the respective polynomials.

2. They are orthogonal on some interval $(a, b) \times (a, b)$ with respect to some measure $W(x, y) = w(x) \otimes w(y)$ in the sense that

$$\iint_a^b \mathrm{d}x\mathrm{d}y \, \phi_i(x)\phi_k(x)\phi_j(y)\phi_\ell(y)w(x)w(y) = \delta_{ik}\delta_{j\ell} \tag{1.5}$$

3. As such, they form a complete basis of all square-integrable $L^2(\mathbb{R}^2)$ functions on the interval $(a, b) \times (a, b)$, that is, if $f(x, y)$ be such a function, then

$$f(x, y) = \sum_{k,\ell=0}^{\infty} C_{k\ell} \, \phi_k(x)\phi_\ell(x)w(x)w(y) \tag{1.6}$$

with constant coefficients $C_{k\ell}$, which are determined by multiplying from the left with $\phi_i(x)\phi_j(y)$ and integrating over all x and y.

$$\iint_a^b \mathrm{d}x\mathrm{d}y \, \phi_i(x)\phi_j(y)f(x, y) = \sum_{k,\ell=0}^{\infty} C_{k\ell} \iint_a^b \mathrm{d}x\mathrm{d}y \, \phi_i(x)\phi_k(x)\phi_j(y)\phi_\ell(y)w(x)w(y)$$

Orthogonality then implies:

$$C_{ij} = \iint_a^b \mathrm{d}x\mathrm{d}y \, \phi_i(x)\phi_j(y)f(x, y) \tag{1.7}$$

4. We can approximate $f(x, y)$ to any desired accuracy by the truncated series

$$f(x) \approx \sum_{k,\ell=0}^{K,L} C_{k\ell} \, \phi_k(x)\phi_\ell(y)w(x)w(y) \tag{1.8}$$

with $K, L < \infty$ and we can make this approximation systematically better by increasing K and L.

Chapter 2

Probability Densities

2.1 The univariate case

If, in particular, we wish to expand a univariate probability density function $p(x)$ with a support of (a, b) into a truncated series of the type shown in eq. 1.4, care has to be taken that key properties of $p(x)$ are conserved.

$$p(x) \geq 0 \tag{2.1}$$

$$\int_a^b \mathrm{d}x\, p(x) = 1 \tag{2.2}$$

One way to satisfy constraint 2.1 is to not directly expand $p(x)$ but, rather, its square root.

$$\sqrt{p(x)} = \sum_{j=0}^{J} c_j\, \phi_j(x) \sqrt{w(x)} \tag{2.3}$$

The probability density is then recovered through $p(x) = \left(\sqrt{p(x)}\right)^2$. Substituting expansion 2.3 into constraint 2.2 and using eq. 1.1 then yields the following requirement for the expansion coefficients.

$$\sum_{i,j=0}^{J} c_i c_j \int_a^b \mathrm{d}x\, \phi_i(x)\phi_j(x)w(x) = \sum_{j=0}^{J} c_j^2 = 1 \tag{2.4}$$

If, on the other hand, we multiply eq. 2.3 with $\phi_i(x)$ and $\sqrt{w(x)}$ and integrate over all x, then we find

$$\begin{aligned}
\int_a^b \mathrm{d}x\, \phi_i(x)\sqrt{w(x)}\sqrt{p(x)} &= \sum_{j=0}^{J} c_j \int_a^b \mathrm{d}x\, \phi_i(x)\phi_j(x)w(x) \\
\int_a^b \mathrm{d}x\, \frac{\phi_i(x)\sqrt{w(x)}}{\sqrt{p(x)}} p(x) &= \sum_{j=0}^{J} c_j \delta_{ij} \\
\mathbb{E}\left[\frac{\phi_i(x)}{\sum_{j=0}^{J} c_j\, \phi_j(x)}\right] &= c_i
\end{aligned} \tag{2.5}$$

where we have exploited eqs. 1.1 and 2.3 again.

2.1.1 Density gradient

For the sake of completeness, we note here that the gradient of the density 2.3 is given by:

$$\frac{\partial}{\partial x} p(x) = 2\left[\sum_{j=0}^{J} c_j \phi_j(x)\right]\left[\sum_{j=0}^{J} c_j \frac{\partial \phi_j(x)}{\partial x}\right] \tag{2.6}$$

2.2 The bivariate case

Equivalently, in two dimensions, we have to demand of the probability density $p(x, y)$ with support $(a, b) \times (a, b)$that:

$$p(x, y) \geq 0 \tag{2.7}$$

$$\iint_a^b \mathrm{d}x\mathrm{d}y\, p(x, y) = 1 \tag{2.8}$$

Again, the former can be satisfied by expanding the square root of $p(x, y)$.

$$\sqrt{p(x, y)} = \sum_{k,\ell=0}^{K,L} C_{k\ell}\, \phi_k(x)\phi_\ell(y)\sqrt{w(x)}\sqrt{w(y)} \tag{2.9}$$

With this expansion, the second demand then translates into a requirement for the expansion coefficients.

$$\sum_{i,j=0}^{K,L} \sum_{k,\ell=0}^{K,L} C_{ij}C_{k\ell} \iint_a^b \mathrm{d}x\mathrm{d}y\, \phi_i(x)\phi_k(x)\phi_j(y)\phi_\ell(y)w(x)w(y) = \sum_{k,\ell=0}^{K,L} C_{k\ell}^2 = 1 \tag{2.10}$$

2.2.1 Density gradient

For the sake of completeness, we give the gradient of the bivariate density as well.

$$\frac{\partial}{\partial x}p(x,y) = 2\left[\sum_{k,\ell=0}^{K,L} C_{k\ell}\phi_k(x)\phi_\ell(y)\right]\left[\sum_{k,\ell=0}^{K,L} C_{k\ell}\frac{\partial\phi_k(x)}{\partial x}\phi_\ell(y)\right] \tag{2.11}$$

$$\frac{\partial}{\partial y}p(x,y) = 2\left[\sum_{k,\ell=0}^{K,L} C_{k\ell}\phi_k(x)\phi_\ell(y)\right]\left[\sum_{k,\ell=0}^{K,L} C_{k\ell}\phi_k(x)\frac{\partial\phi_\ell(y)}{\partial y}\right]$$

Chapter 3

Maximum-Likelihood Estimate of Coefficients

3.1 The univariate case

Say we have observed N data points $x_{n=0\ldots N-1}$, collectively denoted by the vector $\mathbf{x} = \{x_n\}$ and we wish to find an approximation $\hat{p}(x)$ to the underlying probability distribution from which the data points were drawn.

One possibility is to employ the series expansion 2.3 and to determine its coefficients $\mathbf{c} = \{c_j\}$ by maximizing their likelihood $p(\mathbf{x}|\mathbf{c})$ or, rather, by *minimizing* their negative log-likelihood, $-\log p(\mathbf{x}|\mathbf{c})$, under the constraint 2.4. Assuming the data points to be independent and identically distributed, the former is given by

$$\begin{aligned} -\log p(\mathbf{x}|\mathbf{c}) &= -\log \prod_{n=0}^{N-1} p(x_n) = -\log \prod_{n=0}^{N-1} \left(\sqrt{p(x_n)}\right)^2 \\ &= -\sum_{n=0}^{N-1} \log \left(\sum_{j=0}^{J} c_j \, \phi_j(x_n) \sqrt{w(x_n)} \right)^2 \end{aligned} \tag{3.1}$$

and the latter, expressed in terms of a function,

$$q(\mathbf{c}) = \sum_{j=0}^{J} c_j^2 - 1 = 0 \tag{3.2}$$

can be incorporated through a *Lagrange* parameter λ to yield and overall objective function (or Lagrangian $\mathcal{L}$) that needs to be minimized.

$$\mathcal{L}(\mathbf{c}, \lambda|\mathbf{x}) = -\log p(\mathbf{x}|\mathbf{c}) + \lambda q(\mathbf{c}) \tag{3.3}$$

Before we do that, however, it is worthwhile to inspect the properties of a particular quantity.

3.1.1 The Fisher information matrix

The elements $\mathcal{F}_{uv}$ of the *Fisher information matrix* $\boldsymbol{\mathcal{F}}$ for a probability distribution $p(x|\mathbf{c})$ with support (a, b), parameterized by parameters $\mathbf{c}$, can be defined as:

$$\mathcal{F}_{uv} = \int_a^b \mathrm{d}x\, p(x|\mathbf{c}) \frac{\partial}{\partial c_u} \log p(x|\mathbf{c}) \frac{\partial}{\partial c_v} \log p(x|\mathbf{c}) \tag{3.4}$$

Using our $p(x|\mathbf{c}) = \left(\sqrt{p(x|\mathbf{c})}\right)^2$ yields

$$\begin{aligned}
\mathcal{F}_{uv} &= \int_a^b \mathrm{d}x \, \left(\sqrt{p(x|\mathbf{c})}\right)^2 \frac{\partial}{\partial c_u} \log \left(\sqrt{p(x|\mathbf{c})}\right)^2 \frac{\partial}{\partial c_v} \log \left(\sqrt{p(x|\mathbf{c})}\right)^2 \\
&= 4 \int_a^b \mathrm{d}x \, \left(\sqrt{p(x|\mathbf{c})}\right)^2 \frac{\partial}{\partial c_u} \log \left(\sqrt{p(x|\mathbf{c})}\right) \frac{\partial}{\partial c_v} \log \left(\sqrt{p(x|\mathbf{c})}\right) \\
&= 4 \int_a^b \mathrm{d}x \, \left(\sqrt{p(x|\mathbf{c})}\right)^2 \frac{\frac{\partial}{\partial c_u}\sqrt{p(x|\mathbf{c})}}{\sqrt{p(x|\mathbf{c})}} \frac{\frac{\partial}{\partial c_v}\sqrt{p(x|\mathbf{c})}}{\sqrt{p(x|\mathbf{c})}} \\
&= 4 \int_a^b \mathrm{d}x \, \frac{\partial}{\partial c_u}\sqrt{p(x|\mathbf{c})} \frac{\partial}{\partial c_v}\sqrt{p(x|\mathbf{c})}
\end{aligned}$$

and, by substituting expansion 2.3 with derivative

$$\frac{\partial}{\partial c_u} \sum_{j=0}^{J} c_j \, \phi_j(x) \sqrt{w(x)} = \phi_u(x) \sqrt{w(x)} \tag{3.5}$$

and using eq. 1.1, we arrive at

$$\mathcal{F}_{uv} = 4 \int_a^b \mathrm{d}x \, \phi_u(x) \phi_v(x) w(x) = 4\delta_{uv}$$

or, expressed in matrix notation,

$$\boldsymbol{\mathcal{F}} = 4\mathbb{1} \tag{3.6}$$

with $\mathbb{1}$ the unit matrix.

An alternative but equivalent definition of the Fisher information matrix is *via* the expectation value $\mathbb{E}$ of the Hessian $\boldsymbol{\mathcal{H}}$ of the negative log-likelihood,

$$\boldsymbol{\mathcal{F}} = \mathbb{E}\left[\boldsymbol{\mathcal{H}}\right] = -\mathbb{E}\left[\nabla\nabla^{\mathrm{T}} \log p(x|\mathbf{c})\right] \tag{3.7}$$

where ∇ denotes the gradient operator with respect to the parameters $\mathbf{c}$. In our particular case, however, where we have to solve a constrained optimization problem, we have to use the Hessian $\boldsymbol{\mathcal{H}}_{\mathcal{L}}$ of the Lagrangian 3.3,

$$\boldsymbol{\mathcal{H}}_{\mathcal{L}} = \boldsymbol{\mathcal{H}} + \lambda \boldsymbol{\mathcal{H}}_q \tag{3.8}$$

where the second term $\boldsymbol{\mathcal{H}}_q$ is the Hessian of the constraint $q(\mathbf{c})$. In order to evaluate 3.7 with 3.8, we compute individual elements for both matrices, starting

with the first derivative of the negative log-likelihood with respect to coefficient c_v.

$$\begin{aligned}
-\frac{\partial}{\partial c_u}\log p(x|\mathbf{c}) &= -\frac{\partial}{\partial c_u}\log\left(\sqrt{p(x)}\right)^2 \\
&= -2\frac{\partial}{\partial c_u}\log\sqrt{p(x)} \\
&= -2\left(\sqrt{p(x)}\right)^{-1}\frac{\partial}{\partial c_u}\sum_{j=0}^{\infty} c_j\phi_j(x)\sqrt{w(x)} \\
&= -2\left(\sqrt{p(x)}\right)^{-1}\phi_u(x)\sqrt{w(x)}
\end{aligned} \tag{3.9}$$

where we have used eq. 3.5. Now taking the second derivative with respect to c_v, using eqs. 2.3 and 3.1, we find

$$\begin{aligned}
-\frac{\partial^2}{\partial c_u \partial c_v}\log p(x|\mathbf{c}) &= -2\phi_u(x)\sqrt{w(x)}\frac{\partial}{\partial c_v}\left(\sqrt{p(x)}\right)^{-1} \\
&= +2\frac{\phi_u(x)\sqrt{w(x)}\frac{\partial}{\partial c_v}\sum_{j=0}^{\infty} c_j\phi_j(x)\sqrt{w(x)}}{\left(\sqrt{p(x)}\right)^2} \\
&= 2\frac{\phi_u(x)\phi_v(x)w(x)}{p(x)}
\end{aligned} \tag{3.10}$$

and taking the expectation value yields

$$\begin{aligned}
2\,\mathbb{E}\left[\frac{\phi_u(x)\phi_v(x)w(x)}{p(x)}\right] &= 2\int_a^b \mathrm{d}x\, p(x)\frac{\phi_u(x)\phi_v(x)w(x)}{p(x)} \\
&= 2\int_a^b \mathrm{d}x\, \phi_u(x)\phi_v(x)w(x) \\
&= 2\delta_{uv} = 2\mathbb{1}
\end{aligned} \tag{3.11}$$

where we have exploited eq. 1.1 in the last step. Derivatives of the negative log-likelihood with respect to the Lagrange multiplier λ vanish.

Moving on to the second term in 3.8, we derive once with respect to c_u

$$\frac{\partial}{\partial c_u}\lambda q(\mathbf{c}) = \lambda\sum_{j=0}^{\infty}\frac{\partial}{\partial c_u}c_j^2 = 2\lambda c_u = 2\lambda\mathbf{c} \tag{3.12}$$

and once with respect to λ:

$$\frac{\partial}{\partial\lambda}\lambda q(\mathbf{c}) = q(\mathbf{c}) = \sum_{j=0}^{\infty} c_j^2 - 1 = \mathbf{c}^{\mathrm{T}}\mathbf{c} - 1 \tag{3.13}$$

Taking again the derivative of eq. 3.12 with respect to c_v, we find

$$\frac{\partial}{\partial c_v}2\lambda c_u = 2\lambda\delta_{uv} = 2\lambda\mathbb{1} \tag{3.14}$$

and with respect to λ:

$$\frac{\partial}{\partial\lambda} 2\lambda c_u = 2c_u = 2\mathbf{c} \tag{3.15}$$

Finally, we derive eq. 3.13 with respect to c_v to find

$$\frac{\partial}{\partial c_v}\left(\sum_{j=0}^{\infty} c_j^2 - 1\right) = 2c_v = 2\mathbf{c} \tag{3.16}$$

while its derivative with respect to λ vanishes. Taking the expectation value does not change the result. To summarize, if we (arbitrarily) take λ to be the first of parameter, the Hessian 3.8 can be written as:

$$\boldsymbol{\mathcal{H}}_{\mathcal{L}} = \begin{pmatrix} 0 & 2\mathbf{c} \\ 2\mathbf{c} & 2\mathbb{1} + 2\lambda\mathbb{1} \end{pmatrix} \tag{3.17}$$

Comparison to eq. 3.6 then suggest that we would have to require $\lambda = 1$ at the optimal solution point $\mathbf{c}^*$ such that:

$$\boldsymbol{\mathcal{H}}_{\mathcal{L}} = \begin{pmatrix} 0 & 2\mathbf{c} \\ 2\mathbf{c} & \boldsymbol{\mathcal{F}} \end{pmatrix} \tag{3.18}$$

As we will see below, this is certainly the case if, as we have just done, we compute the likelihood at a single point x only.

3.1.2 Explicit expressions

Regardless of how we chose to minimize the objective function 3.3 in practice, it might be useful to know its gradient and, potentially, also its Hessian. In analogy to eq. 3.9, we find for the gradient ∇_{nll} of the negative log-likelihood 3.1 with respect to the coefficients

$$\nabla_{nll} = -2\sum_{n=0}^{N-1} \frac{\boldsymbol{\Phi}(x_n)\sqrt{w(x_n)}}{\sqrt{p(x_n)}} = -2\sum_{n=0}^{N-1} \frac{\boldsymbol{\Phi}(x_n)}{\mathbf{c}\boldsymbol{\Phi}(x_n)} \tag{3.19}$$

where $\boldsymbol{\Phi}$ stands for the vector $\{\phi_i\}$, the denominator is written as a dot product, and $\sqrt{p(x_n)}$ was substituted by its expansion 2.3. Given result 3.12, deriving the coefficient gradient of the constraint 3.2 is straightforward.

$$\nabla q(\mathbf{c}) = 2\mathbf{c} \tag{3.20}$$

Together with eqs. 3.1 and 3.2, these gradients could already be used to perform a *constrained* minimization of the negative log likelihood. If, however, we wish to perform a direct minimization of the objective function 3.3, we also need to develop an expression for the Lagrange parameter λ. At the optimal solution point $\mathbf{c}^*$, we require

$$\nabla\mathcal{L} = \nabla_{nll} + \lambda^*\nabla q = 0$$

or, using, eqs. 3.19 and 3.20:

$$-2\sum_{n=0}^{N-1}\frac{\boldsymbol{\Phi}(x_n)}{\mathbf{c}^*\boldsymbol{\Phi}(x_n)}+2\lambda^*\mathbf{c}^*=0$$

$$\mathbf{c}^*=\frac{1}{\lambda^*}\sum_{n=0}^{N-1}\frac{\boldsymbol{\Phi}(x_n)}{\mathbf{c}^*\boldsymbol{\Phi}(x_n)} \tag{3.21}$$

Comparison with eq. 2.5 then shows that

$$\lambda^*=N \tag{3.22}$$

for the right-hand side to be an unbiased estimate of the expectation value found there for the optimal coefficients. Following eq. 3.13, the λ-component of the gradient is simply the (square-)normalization condition of the coefficients. Should we require also the Hessian in our optimization scheme we employ eq. 3.17, potentially even as just a constant matrix with λ set to its optimal value given by eq. 3.22.