Why P-splines?

Paul Eilers & Brian Marx

June 7, 2021

1 Introduction

P-splines are the ultimate smoothing tool. You can use them for straightforward smoothing like computing a trend in a scatter plot or estimating a probability density. They can also be used in more complicated application, like varying-coefficient models, life tables, and spatial models, among others. This document has two goals: one is to give a (very) short technical background; the other is to explain what P-splines are such a good tool. We also discuss a number of misconceptions that we have encountered in the literature.

You may have found this document on our website (`psplines.bitbucket.io`), dedicated to the book *Practical Smoothing. The Joys of P-splines* (Eilers and Marx, 2021). If not, we suggest that you have a look on that site. The book provides a detailed description of P-splines and it shows them in action on a rich variety of practical applications. Each and every example is illustrated by a figure graph and for each figure code is supplied for reproducing it exactly. There is also a package, `JOPS`, that provides useful functions and example data.

To use P-splines, some choices have to be made: number and degree of the B-splines, their domain, the order of the penalty, and the value of the smoothing parameter. Here we systematically work through them and indicate sensible default values for most of them. The functions in the `JOPS` package use these defaults.

2 P-splines illustrated

Figure 1 shows in a graph the core idea of P-splines, as applied to smoothing of a scatter plot. The small grey dots show simulated data. They have been connected by thin grey lines for clarity. The thick blue line is the computed P-spline trend. It is the sum of the rainbow-colored "mountains" below it. They are B-splines, scaled by coefficients. The values of the coefficients are shown by the large colored dots.

It is easy to see that to get an optimal fit (in the least squares sense) of the trend of the data, one could use linear regression. For each B-spline, its values at all observed x are collected in a separate column of a matrix, say B. Minimizing $||y - B\alpha||^2$ gives $\hat{\alpha}$ and $B\hat{\alpha}$ gives the best fit. Once $\hat{\alpha}$ is available, we can compute a fitted curve with any desired resolution, as $\tilde{B}\hat{\alpha}$, by filling the columns of $\tilde{B}$ with the values of the B-splines computed at that resolution.

Note that all B-splines have the same width. In principle it is possible to have variable widths, but we don't se that option in our work.

We described regression on B-splines; it can be an effective recipe in many cases. The main disadvantage is that the smoothness of the result is basically determined by the number of B-splines. Increasing that number gives a less smooth result. Also, with sparse data, it can happen

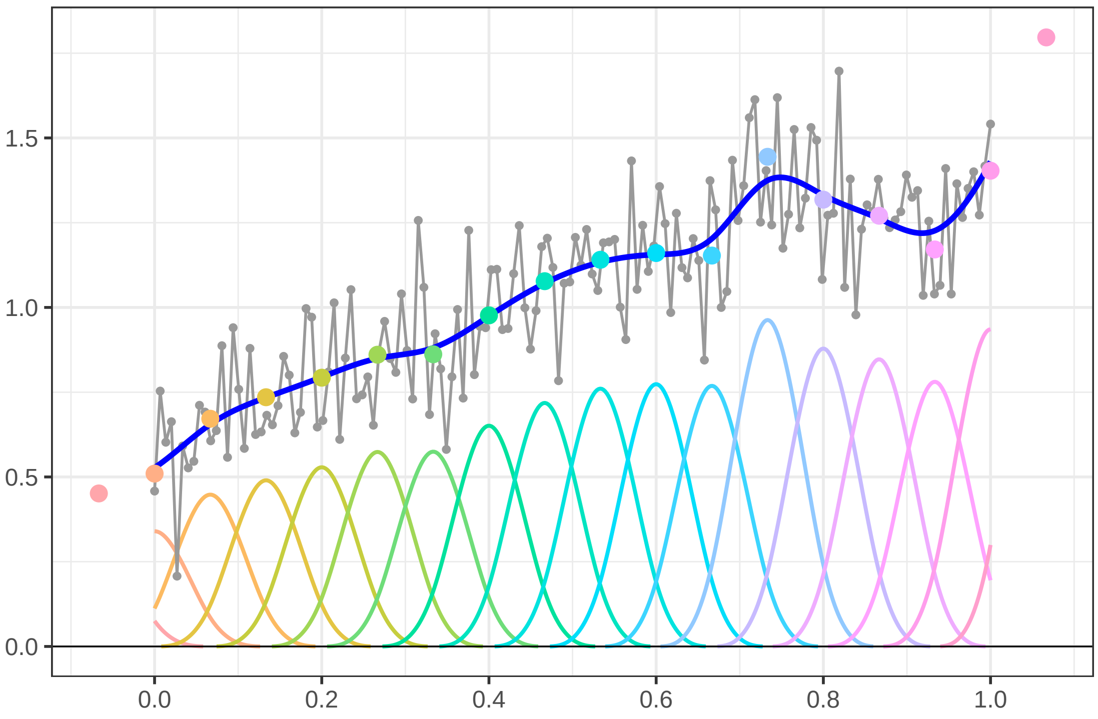

Figure 1: The core idea of P-splines: a sum of B-spline basis functions, with gradually changing heights. The small connected grey dots show simulated data. The blue curve shows the P-spline fit, and the large dots the B-spline coefficients (they have the same colors as the splines).

that some of B-splines have no support, because of missing x "under" those B-splines. One or more columns of B will be empty then, and the regression will not work.

P-splines add a penalty to tune smoothness continuously and to eliminate problems with missing support. It is a simple penalty: restrain the differences between neighboring elements of the coefficient vector α. In its simplest form the penalty is $\lambda \sum_j (\alpha_j - \alpha_{j-1})^2$. The objective function is

$$||y - B\alpha||^2 + \lambda \sum_j (\alpha_j - \alpha_{j-1})^2,$$

which can be written compactly as

$$||y - B\alpha||^2 + \lambda ||D\alpha||^2.$$

The parameter λ tunes the penalty: increasing its value gives a smoother result. Here D is a matrix such that $D\alpha$ forms differences of α. The explicit solution is

$$\hat{\alpha} = (B'B + \lambda D'D)^{-1} B'y.$$

Modern languages like R and Matlab have built-in provisions for matrix operations, so it takes only a few lines of computer code to apply P-splines.

In the example, D is based on first order differences. But we have complete freedom to use higher order differences, or any difference equation to determine D. In Chapter 8 of our book we call this "designer penalties." They are great for periodic or adaptive smoothing, among many other possibilities.

Figure 2 illustrates the influence of the penalty, when smoothing the same data with the same set of B-splines, while varying λ.

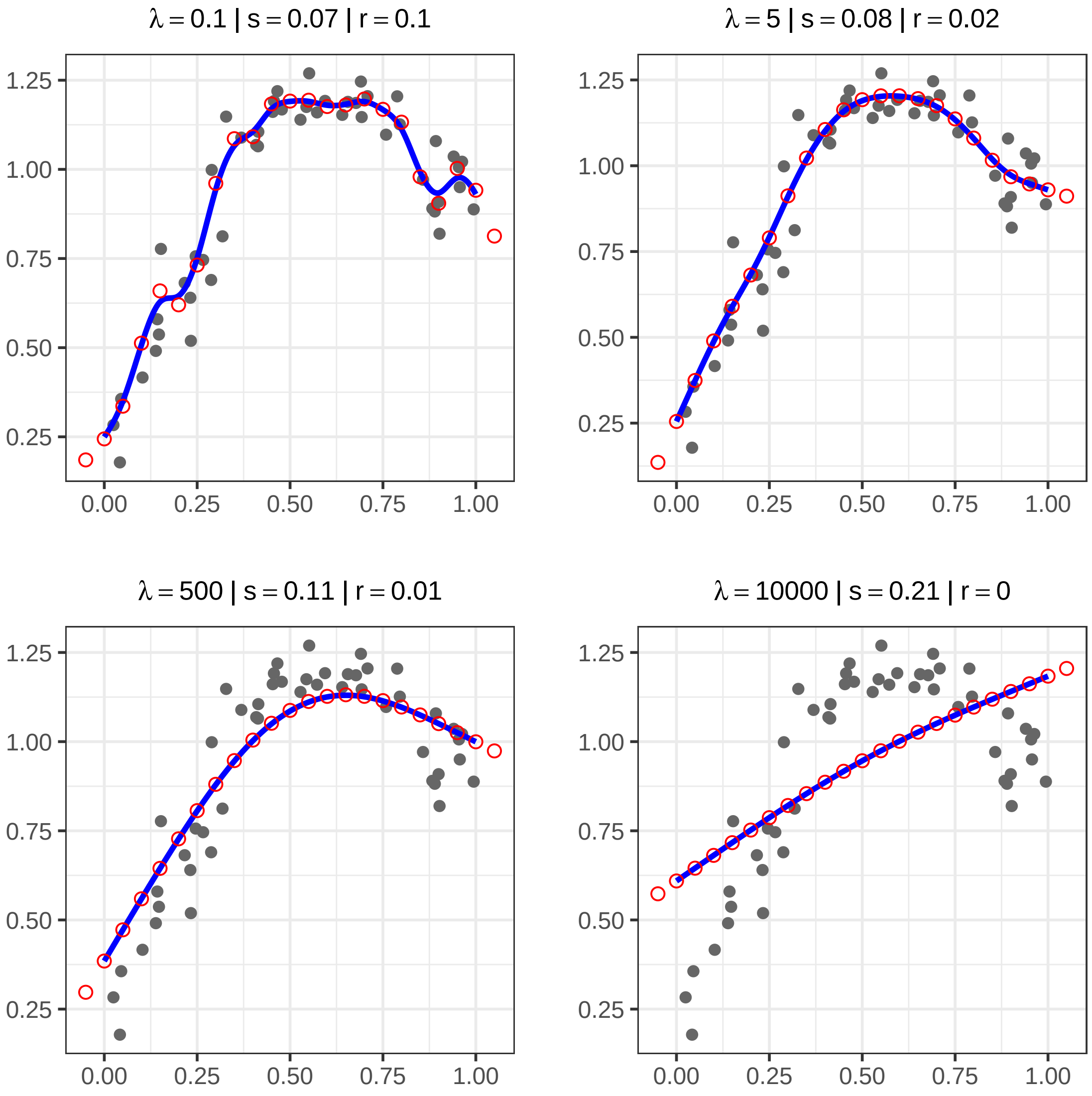

Figure 2: Illustration of the fit of P-splines when varying the strength of the penalty (parameter λ). The number and the widths of the B-splines do not change, but their coefficients (the red circles) are more constrained when λ is increased.

3 P-splines parameters

To use P-splines, values have to be chosen for five parameters. We do not give many details, as they can be found in our book.

The domain. This is the region of the x axis on which the B-splines are defined. Often the minimum and maximum of x can be chosen as left and right boundaries (the default choice). In other cases, one can round them to pleasant numbers. It can do no harm to choose the domain (much) wider than the range of x. Extrapolation occurs automatically.

The domain should be wide enough to include all observed x. If this is not the case, the function bbase (that computes a B-spline basis matrix) automatically extends the domain to min(x) at the left and max(x) at the right. It also produces a warning if this happens.

The degree of the B-splines. The default choice is 3, giving cubic B-splines that consist of smoothly joining cubic polynomial segments. In practice there is seldom a need for another value.

The number of B-splines The default choice is 10, but more is a safe bet. For some data, a very flexible curve may be needed and 10 B-splines might not be enough. It is *impossible* to have too many B-splines, because the penalty removes any singularities.

The order of the penalty. The default choice is 2. Generally it strikes a good balance between smoothness of the fit and closeness to the data. In special cases a first or third order penalty is advisable.

The smoothing parameter. This is λ, the key element of P-splines, and there is no default choice. If your data are well behaved, a good value can be determined automatically through a variety of methods, but there is no guarantee that they will always give a meaningful result. For instance, smooth trends with serially correlated errors can lead to surprises, i.e. much less smoothing than expected.

It always is a good strategy to explore a range of values for λ and judge the results visually. Such a range should be large and use a linear grid of values for $\log_{10}(\lambda)$.

In most cases, two basis matrices are computed: one for fitting the data and the other for plotting the fitted curve on a nice grid. It is crucial that domain, degree and number of the B-splines in the second matrix are equal to those in the first.

4 P-spline advantages

P-splines combine (relatively many, evenly spaced) B-splines with a discrete roughness penalty on their coefficients. This gives them many practical and theoretical advantages:

- The B-splines all have the same shape and are evenly spaced; optimal knot placement is not an issue.
- Thanks to the penalty, the number of B-splines can be chosen freely. It is not possible to have too many B-splines.
- B-splines of any degree can be computed quickly and easily. In many cases linear B-splines work well. They are extremely easy to compute. With many knots, they give a pleasing piecewise linear fit.
- A B-spline basis matrix is intrinsically sparse. Our software can compute very large B-splines bases in a sparse matrix format efficiently. The penalty matrix is also sparse. Using sparse matrix software, data series with millions of observations can be smoothed in a fraction of as second.
- A P-spline model is *parametric*. The B-spline coefficients are the parameters. They are close to the local function values. They have a direct and clear interpretation. This is not the case for most parametric models. For linear P-splines, the fitted curve is obtained by connecting the dots that represent the coefficients.

- The penalty is the key element. Usually it is based on (higher order) differences of the coefficients. Its order can be chosen freely, independent of the degree of the B-splines. More general difference equations can be used in special cases, as for periodic or circular data.

- The discrete penalty is *not* an approximation to a continuous one. The popular integrated squared second derivative, i.e. the one we know from smoothing splines or O'Sullivan (1986), demands a curve fit that consists of polynomial pieces of degree three or higher; otherwise the penalty disappears.

- P-splines are based on (penalized) regression, so non-normal data can be handled with ease, adapting the generalized linear model framework.

- Numerous extensions are easy to implement, such as additive and varying coefficient models, quantile and expectile smoothing, signal regression, the composite link model, among others.

- P-splines can be interpreted and analysed as mixed models. Penalty parameters become ratios of variances. Fast algorithms can estimate multiple penalty parameters with ease.

- Bayesian P-splines can be realized easily, using Markov chains or Laplace approximation. Either framework computes tuning parameters automatically.

- The effective dimension of a P-spline model is well defined and easy to compute. It is useful for quantifying model complexity, cross-validation, and the computation of AIC.

- Tensor products of B-spline bases and extended penalties generalize P-splines for multidimensional smoothing. Large data sets can be handled straightforwardly. Data on huge grids (1000 by 1000 cells, or larger) are no problem, because array algorithms make the computations highly efficient and fast.

5 Popular myths about P-splines

There exist myths about P-splines that linger on, although we have shown them to be incorrect on several occasions; we comment on them here.

The difference penalty approximates a derivative penalty. This is not true. For a derivative penalty to work, a piecewise-continuous curve fit of high enough degree is needed. Otherwise the derivative vanishes, and the penalty cannot do its work. With a continuous penalty, the order of the penalty and the degree of the B-splines are strongly coupled. A difference penalty on the coefficients of the B-splines does not have this problem. Any degree of the B-splines can be combined with any order of the penalty.

The knots of B-splines should be quantiles of the abscissae. This is only true when no penalty is being used, as otherwise some B-splines may have no support. Always use a penalty, because it eliminates the effects of no support, in addition to smoothing. Evenly spaced knots are a natural and easy choice.

You should use less B-splines than observations. This is another manifestation of the "no support" myth. It is not true. The number of B-spline should only be larger than the order of the penalty. This myth is responsible for the fact that you cannot always compute a rich B-spline basis in R, when using the `bs()` function in the `spline` package. This function simply refuses to do the computation. Our function `bbase()` does not have this problem.

It is a good idea to optimize the number of knots. It is not. It means a lot of work, with minimal rewards. Rules of thumb, like that of Ruppert (2002) have no value.

The smoothing spline is sacred. It is not. Unfortunately, it has been glorified, with the unpleasant consequence that the continuous penalty became the standard. See the first item in this list.

O'Sullivan splines are better. This claim of Wand and Ormerod (2008) is wrong. They use a wrong B-spline basis, one with multiple knots at the ends. See Eilers et al. (2015). O'Sullivan's penalty is based on the integral of a squared (higher) derivative of the fitted curve, so it has the drawbacks that were mentioned in the first item of this list.

References

Eilers, P. H. C., Marx, B. D., and Durbán, M. 2015. Twenty years of P-splines. *SORT-Statistics and Operations Research Transactions*, **39**(2), 149–186.

Eilers, P.H.C, and Marx, B.D. 2021. *Practical Smoothing: The Joys of P-splines.* Cambridge University Press.

O'Sullivan, F. 1986. A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, **1**, 505–527.

Ruppert, D. 2002. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, **11**, 735–757.

Wand, M. P., and Ormerod, J. T. 2008. On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, **50**, 179–198.