A note on the estimation and simulation of distributions with Bernstein polynomials

Dietmar Pfeifer

Carl von Ossietzky Universität Oldenburg, Germany

October 07, 2022 (revised version)

Abstract: We present a study on the estimation and Monte Carlo simulation of continuous distributions, in particular with infinite support, with Bernstein polynomials extending previous approaches in this direction.

Key words: Bernstein polynomials, Monte Carlo simulation, risk management
AMS Classification: 41A10, 11K45, 62G05

1. Introduction

Bernstein polynomials came to light with the pioneering paper by Serge Bernstein in 1912 [1] and has ever since become an indispensable tool in calculus and approximation theory (see e.g. [2]). In this paper, we comment on methods to estimate the quantile function of a continuous distribution – especially with infinite support – using Bernstein polynomials in a subtle way, extending similar ideas of Babu et al. [3] or Vil'chevskii and G. L. Shevlyakov [4]. This allows for an easy way to simulate continuous distributions on the basis of given data, in particular for risk management purposes when the estimation of a risk measure like Value at Risk (VaR) is required.

2. The general setup

Suppose that *n* i.i.d. observations $X_1,\cdots,X_n$ distributed as a risk X are given. We assume that these observations come from a fixed, but unknown continuous distribution P^X with cdf F, concentrated on a – possibly infinite – interval I. One often assumes that P^X belongs to a certain class of distributions like lognormal, Fréchet, Pareto etc. There are several methods to estimate P^X or F, resp., for instance by a Q-Q-plot or other statistical procedures (see e.g. Pfeifer [5]). Denote the estimated cdf by $\hat{F}$. In risk management, one is often interested in larger Monte Carlo simulations for P^X, for instance for the estimation of a risk measure if X is one out of several risks with a certain dependence structure.
We can assume that the X_i can be represented as $X_i = Q(U_i)$ with the quantile function $Q = F^{-1}$ and independent standard uniform random numbers U_i. For a Monte Carlo study, we need to know Q or a good approximation of it, based on the given sample. Now let G be a (first of all) arbitrary continuous and strictly increasing cdf with support I. Then obviously $Q(u) = G^{-1}\big(G(Q(u))\big),\ 0<u<1$. Our idea is the following:

Approximate $G(Q(u))$ by a Bernstein polynomial $B(u)$ in an appropriate way on the basis of the given sample and use $G^{-1}(B(u)),\ 0<u<1$ as approximate quantile function for the risk X. Note that $G(Q(u))$ is bounded with $G(Q(0))=0$ and $G(Q(1))=1$.

From a statistical point of view, it might be wise to use an empirical estimate G for the true underlying cdf F. Then the procedure is as follows:

1. Transform the observations according to $Y_k := G\left(X_{k:n}\right),\ k = 1,\cdots,n$ where $X_{k:n}$ denotes the k-th order statistic, and put $Y_0 := 0,\ Y_{n+1} := 1$.
2. Calculate the corresponding (random) Bernstein polynomial, i.e.

$$B_n(u) := \sum_{k=0}^{n+1} \binom{n+1}{k} Y_k\, u^k (1-u)^{n+1-k}, u = 0,\cdots,1$$

   and use $G^{-1}\left(B_n\right)$ as an approximation for the true underlying quantile function F^{-1}.

The advantage of this approach is that due to the boundedness of the $Y_k = G\left(X_{k:n}\right)$ we get no problems with the tails, other than when Bernstein polynomials are directly used as an interpolation of the empirical distribution function.

Example. We consider the data given in Cottin and Pfeifer [6] with $n = 20$. The fist risk is assumed to be lognormally distributed, or, alternatively, the log data X_i are assumed to be normally distributed. We use the estimated location and scale parameters $\hat{\mu}$ (empirical mean of log data) and $\hat{\sigma}$ (empirical standard deviation of log data) for a normal cdf G with these parameters as in Cottin and Pfeifer [6].

Table 1. Data and their transformations for the first risk in the above Example.

k	1	2	3	4	5	6	7	8	9	10
$X_{k:n}$	–2.765	–1.483	–0.853	–0.759	–0.392	–0.200	–0.194	–0.144	–0.041	0.169
$G\left(X_{k:n}\right)$	0.005	0.076	0.195	0.219	0.330	0.395	0.397	0.414	0.451	0.527

k	11	12	13	14	15	16	17	18	19	20
$X_{k:n}$	0.182	0.247	0.351	0.438	0.666	0.679	0.713	1.088	1.907	2.298
$G\left(X_{k:n}\right)$	0.531	0.555	0.592	0.622	0.697	0.701	0.712	0.816	0.950	0.977

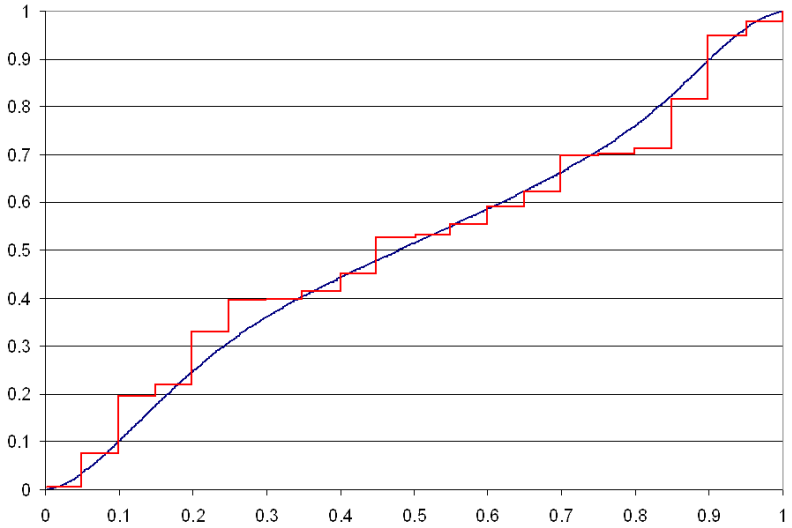

Figure 1. Graph of *G*-transformed empirical quantile function (red) and corresponding Bernstein polynomial (blue)

With this approach, we get an approximate Value of Risk VaR_α for a risk level of $\alpha = 0.005$ of $\mathrm{VaR}_\alpha = 24.558$ while with the estimated lognormal distribution from [6], we only get $\mathrm{VaR}_\alpha = 18.911$.

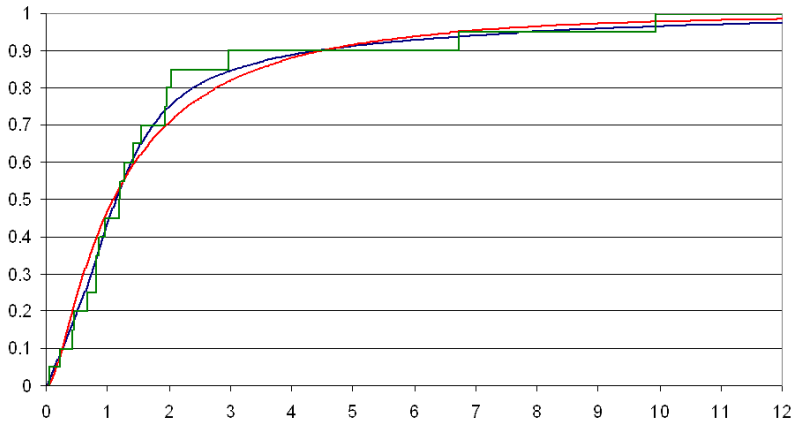

Figure 2. Graph of empirical cdf (green), estimated lognormal cdf (red) and *G*-transformed cdf (blue)

As can clearly be seen, the *G*-transformed cdf is closer to the empirical cdf than the statistically estimated lognormal cdf.

The second risk is assumed to be Fréchet distributed, or, alternatively, the log data Y_i are assumed to be Gumbel distributed. We use the estimated location and scale parameters $\hat{\mu}$ and $\hat{\sigma}$ as $\hat{\mu} = m - \gamma\frac{\sqrt{6}}{\pi}s$ and $\hat{\sigma} = \frac{\sqrt{6}}{\pi}s$ with Euler's constant $\gamma = 0.577216...$ and the empirical mean m and the empirical standard deviation s for a Gumbel cdf G with these parameters.

Table 2. Data and their transformations for the second risk in the above Example.

k	1	2	3	4	5	6	7	8	9	10
$Y_{k:n}$	–0.342	–0.178	–0.112	–0.109	–0.106	–0.086	–0.069	–0.045	–0.035	0.008
$G(Y_{k:n})$	0.039	0.182	0.268	0.272	0.275	0.305	0.328	0.363	0.378	0.439

k	11	12	13	14	15	16	17	18	19	20
$Y_{k:n}$	0.040	0.063	0.074	0.112	0.132	0.150	0.290	0.535	0.810	0.985
$G(Y_{k:n})$	0.484	0.515	0.530	0.577	0.602	0.624	0.761	0.901	0.965	0.982

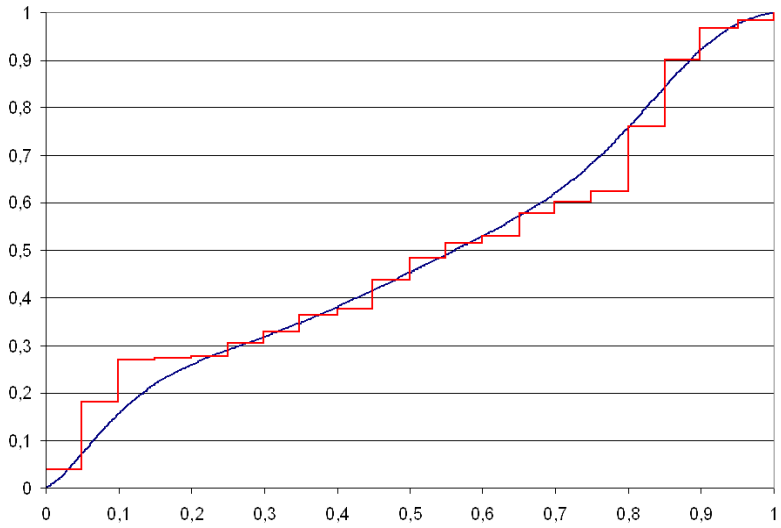

Figure 3. Graph of G-transformed empirical quantile function (red) and corresponding Bernstein polynomial (blue)

With this approach, we get an approximate Value of Risk VaR_α for a risk level of $\alpha = 0.005$ of $\mathrm{VaR}_\alpha = 4.770$ while with the estimated Fréchet distribution, we get only $\mathrm{VaR}_\alpha = 3.708$.

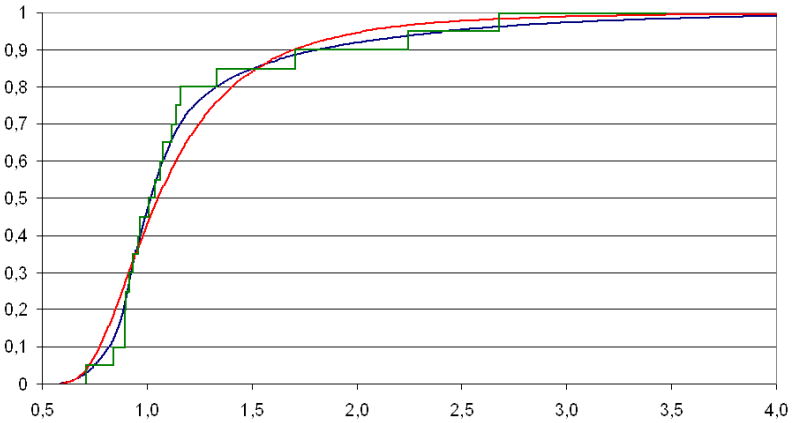

Figure 4. Graph of empirical cdf (green), estimated Fréchet cdf (red) and G-transformed cdf (blue)

As can clearly be seen, the *G*-transformed cdf is again closer to the empirical cdf than the estimated Fréchet cdf.

Finally, we consider combined ratio data from a buildings insurance (VGV) with $n = 18$, taken from [5]. For a first approximation, we use a lognormal distribution, i.e. a normal distribution for the log data. The estimated location and scale parameters are $\hat{\mu} = 0.0781$ (empirical mean of log data) and $\hat{\sigma} = 0.1180$ (empirical standard deviation of log data).

Table 3. Log data and their transformations for the VGV risk in the above Example.

k	1	2	3	4	5	6	7	8	9
$X_{k:n}$	-0,041	-0,024	-0,021	0,002	0,014	0,018	0,019	0,028	0,033
$G(X_{k:n})$	0,157	0,193	0,200	0,259	0,293	0,305	0,308	0,334	0,352

k	10	11	12	13	14	15	16	17	18
$X_{k:n}$	0,037	0,042	0,070	0,070	0,091	0,115	0,291	0,298	0,365
$G(X_{k:n})$	0,365	0,380	0,471	0,474	0,543	0,623	0,965	0,969	0,992

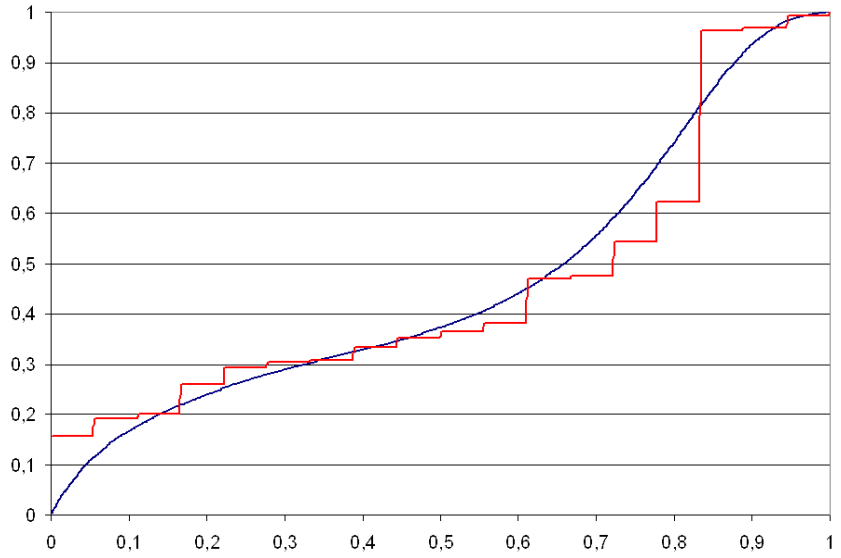

Figure 5. Graph of *G*-transformed empirical quantile function (red) and corresponding Bernstein polynomial (blue)

With this approach, we get an approximate Value of Risk VaR_α for a risk level of $\alpha = 0.005$ of $\widehat{\mathrm{VaR}}_\alpha = 157.05\%$ while with the estimated lognormal distribution from [5], we only get $\widehat{\mathrm{VaR}}_\alpha = 146,51\%$.

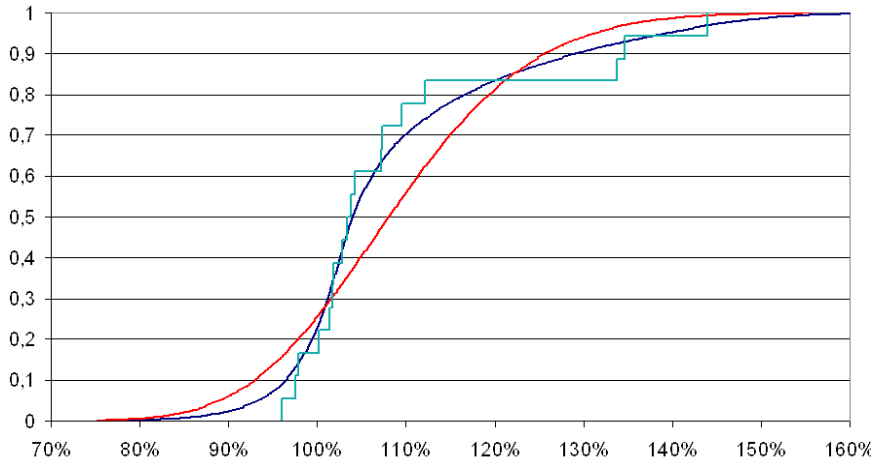

Figure 6. Graph of empirical cdf (green), estimated lognormal cdf (red) and *G*-transformed cdf (blue)

As can clearly be seen, the *G*-transformed cdf is again closer to the empirical cdf than the statistically estimated lognormal cdf.

3. Consistency

Note that the $Y_k := G\left(X_{k:n}\right),\ k = 1,\cdots,n$ can be considered as order statistics $Z_{k:n}$ from random variables $Z_k := G(X_k),\ k = 1,\cdots,n.$ These are bounded, hence their quantile function $Q_Z(u) = G\left(F^{-1}(u)\right),\ 0 < u < 1$ is uniformly continuous. This means that the convergence of the empirical quantile functions pertaining to the $Z_{k:n} = G(X_{k:n})$ is uniform to Q_Z, cf. prop. 5, p. 250 in Fristedt and Gray [7], the same being valid for the completed empirical quantile function by adding the values $Z_0 := 0,\ Z_{n+1} := 1.$ Hence the random Bernstein polynomial converges almost surely (a.s.) with limit Q_Z and hence $G^{-1}\left(B_n\right)$ converges a.s. to F^{-1} with increasing sample size, which implies consistency of the method proposed.

4. Conclusions

Estimating underlying risk distributions and their Monte Carlo simulation is an important task in risk management which sometimes leads to an underestimation of the true risk measures if only standard statistical methods are used. The approach which we suggest in this paper overcomes the problem of estimating an appropriate tail behaviour of the risk distributions, in particular if the underlying support is unbounded. Practical examples from the insurance industry show that our method often gives higher estimates for risk measures than with standard statistical methods, which is probably a desirable fact for a cautious estimation of the overall risk of an insurance company especially under regulatory requirements.

References:

1. Bernstein, S. Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités. *Commun. Kharkov Math. Soc.* **1912**, 13, 1–2.
2. Lorentz, G.G. *Bernstein Polynomials*, 2nd ed.; Chelsea Publishing Company: New York, NY, USA, 1986.
3. Babu, G.J.; Canty, A.J.; Chaubey, Y.P. Application of Bernstein polynomials for smooth estimation of a distribution and density function. *J. Statist. Plann. Inference* 2002, 105, 377–392.
4. Vil'chevskii, N.O.; Shevlyakov, G.L. On the Bernstein polynomial estimators of distribution and quantile functions. *J. Math. Sci.* 2001, *105*, 2626–2629.
5. Pfeifer, D. Modellvalidierung mit Hilfe von Quantil-Quantil-Plots unter Solvency II. *ZVersWiss* (2019) *108*:307–325. https://doi.org/10.1007/s12297-019-00451-y
6. Cottin, C.; Pfeifer, D. From Bernstein polynomials to Bernstein copulas. *J. Appl. Funct. Anal.* 2014, 9, 277–288.
7. B. Fristedt and L. Gray: A Modern Approach to Probability Theory, Birkhäuser, Basel 1997.