Journal of Biometrics & Biostatistics - Open Access

Research Article
OPEN ACCESS Freely available online
doi:10.4172/2155-6180.1000105

Modeling the Log Density Distribution with Orthogonal Polynomials

Eugene Demidenko

Dartmouth Medical School, Section of Biostatistics and Epidemiology, New Hampshire, USA

Abstract

A possibly multimodal log density is modeled via orthogonal polynomials. An efficient Fisher scoring algorithm for maximum likelihood density estimation is described. Statistical hypothesis testing is emphasized such as the test for normality and density multimodality. The density estimation is illustrated with two biomedical examples: brain oxygen distribution and toenail arsenic distribution among New Hampshire residents.

Keywords: Bump hunting; Fisher scoring; Hypothesis testing; Maximum likelihood; Multimodality

Introduction

Density estimation is one of the most important and difficult statistical problems. There exists a vast literature on the topic and this is not the goal of the paper to provide an overview of all existing approaches. Mainly three methodologies have been developed: First, a nonparametric approach can be used, such as kernel density estimation, penalized maximum likelihood or spline density smoothing [12]. Second, finite mixture can be applied to model multimodal distributions [13]. Third, polynomials can be used for the log density modelling and estimation. The latter approach, taken in the present paper, is the simplest and can be used at a preliminary stage of density estimation.

The rationale for using polynomials of a higher order for the log density is as follows. First, the log density of the normal distribution is the polynomial of the second order. Second, since polynomials may have several minima and maxima of the higher order they can describe multimodal distributions, as does the nonparametric approach. Third, by testing the statistical significance of the efficients of the polynomial greater than two we can test the hypothesis that the distribution is normal.

We hypothesize that (a) the real data distribution do not have many components, say two or three and (b) the density function is fairly rounded. If so, expressing the density function through a polynomial of order 4 or 6 becomes justifiable. For example, a polynomial of the fourth order may describe a bimodal density pointing out to the presence of a subpopulation with outstanding values. As was mentioned above, the advantage of polynomial density is that it is parametric and therefore classic statistical hypothesis testing applies. For example, to test that the density is Gaussian, we simply test the hypothesis that the coefficients at the powers higher than 2 are zero; to test that the density is unimodal, we test that the polynomial derivative has a unique root. Especially attractive, in the framework of exponential polynomial fitting, is statistical hypothesis testing such as "Is density Gaussian or bimodal?"

Several authors, including [1] and [2] used polynomials to model density. Although this approach is computationally straightforward it may produce negative values of the density estimate. Following [3] and [9], we use exponential polynomials that guarantees positive density values. An important argument in favor of modeling density via exponential polynomial is the fact that if the order of polynomial is 2 one obtains a normal distribution. Unlike previous authors, we do not use penalization since we advocate a moderate polynomial order. Since our approach is completely parametric, we apply statistical hypothesis testing to chose the order of the polynomial. The goal of the present paper is to concentrate on computationally effective estimation algorithms with relevant hypothesis testing.

The structure of the papers is as follows. In the next section we describe a parametric model for the log density via an orthogonal polynomials. In section 3, we develop the Fisher Scoring algorithm for maximum likelihood estimation. In section 4, we illustrate density estimation with two examples, the rat brain oxygen distribution and toenail arsenic concentration in New Hampshire residents. We show how to statistically test that the distribution is normal.

Density Model

We model the log density distribution via a polynomial, or equivalently, the density f via an exponential polynomial of order $K \geq 2$ as

$$f(y;\alpha) = \frac{1}{c(\alpha)} e^{-\Sigma_{k=1}^{K} \alpha_k P_k(y)}, \quad -\infty < y < \infty, \tag{1}$$

where $P_k(y)$ is a polynomial of the kth order with known coefficients,

$$c(\alpha) = \int_{-\infty}^{\infty} e^{-\Sigma_{k=1}^{K} \alpha_k P_k(y)} dy \tag{2}$$

is the normalizing coefficient and $\alpha = (\alpha_1, ..., \alpha_K)'$ is the parameter vector (boldface is used for vectors and matrices). Note that there is no constant term ($k > 0$) because it is saturated in the normalizing coefficient. Thus, we assume that observations are continuously distributed on $(-\infty,\infty)$. In a special case $P_k(y) = y^k$; this parametrization will be called *canonical*. Alternatively we may call the density model (1) as exponential polynomial.

Since polynomial values are highly correlated, the problem of the coefficients estimation becomes ill-posed. To facilitate

Corresponding author: Eugene Demidenko, Dartmouth Medical School, Section of Biostatistics and Epidemiology, 7927 Rubin Building, One Medical Center Drive, Lebanon, NH 03756, USA, Tel: (603) 653-3682; E-mail: eugened@dartmouth.edu

Received October 30, 2010; **Accepted** November 09, 2010; **Published** November 12, 2010

Citation: Demidenko E (2010) Modeling the Log Density Distribution with Orthogonal Polynomials. J Biomet Biostat 1:105. doi:10.4172/2155-6180.1000105

computations, orthogonal polynomials have been proposed.There are several systems of orthogonal polynomials on the interval [0,1], for example. Laguerre, Legendre, or Chebyshev, that differ by the definition of the scalar product defined via an integral [11]. Usually, Hermite polynomials are used for density modeling defined on (-∞,∞) with the weight function e^{-x^2}. Importantly, coefficients of orthogonal polynomials from different systems are linearly dependent. For example, if $P_k(y)=\Sigma_{j=1}^{k} a_j y^j$ with coefficients collected in a $k \times 1$ vector, a, and $H_k(y)$ is Hermite polynomial with the $k \times 1$vector of coefficients, b, then there exist a $k \times k$ nonsingular matrix, M, such that b = Ma. Consequently, different polynomial systems simply imply different linear parametrizations of α in the density estimation (1).

We, however, prefer a data-based definition of polynomial orthogonality, namely,

$$\sum_{i=1}^{n} P_j(yi) P_k(yi) = 0 \text{ for } j \neq k,$$

where y_1, $\mathbf{y}_2$, ..., y_n are observations. In fact, we expect that the data-based orthogonality would be more efficient because statistical computations are carried out on the sample values,not in the functional space defined on (-∞,∞) as is assumed for Hermite polynomials. Databased orthogonal polynomials are readily available in popular statistical packages such as S-Plus (function poly).

Exponential polynomial density estimation (1) leads to a nonlinear statistical problem that involves integration (2). When K > 2, there is no explicit formula for c(α) and numerical integration is required–this is the major technical problem in estimating parameters of density (1).

Maximum Likelihood Estimation

If $\{yi, i = 1, 2, ..., n\}$ are iid observations from *f*, the log-likelihood function takes the form

$$l(\alpha) = -n \ln c(\alpha) - \sum_{i=1}^{n} \sum_{k=1}^{K} \alpha_k P_k(yi).$$

In this section, we assume that *K* is an even known number and $\alpha_K > 0$. These conditions imply that the integral (2) exists. In fact, positiveness and statistical significance of the $\alpha'_k S$ estimate may be a guideline for choosing an appropriate *K*.

Without loss of generality one can assume that observations are ascending, namely, $y_1 \leq y_2 \leq ... \leq y_n$. There exists no closed form solution of the integral (2) for $K > 2$, so we apply numerical integration. Generally, the integral $\int_{-\infty}^{\infty} G(y)dy$ is approximated with a sum,

$$\int_{-\infty}^{\infty} G(y)\,dy \approx \sum_{q=1}^{Q} w_q G(z_q), \tag{3}$$

where w_q are the weights and $\{z_q, q = 1, 2, ...,Q\}$ is the sample of the argument in the range where the integrand value is not negligible. The easiest way to derive the weights is to take observations themselves as the sample for numerical integration, we refer to this algorithm as the *y-value* method. Then, from the trapezoid rule, we obtain

$$z_q = y_q, \quad w_q = y_{q+1} - y_q, \quad q = 1,...,Q = n-1. \tag{4}$$

Another way to obtain the sample of *z*-values is to take equidistant values within the interval (min *y* - *K*(max *y* - min *y*), max *y* + *K*(max *y* - min *y*)), where *K* can be chosen 1/2 or 1/3 to cover the range of possible *y*-values. Finally, Gauss-Hermite quadrature [5] applies with $z_q = \overline{y} + \hat{\sigma}\sqrt{2}x_q, w_k = w_k \hat{\sigma}\sqrt{2}e^{x_q^2}$, where x_q and ω_k are Gauss-Hermite nodes and weights (these values can be computed using a C-code gauher [11]).

As a word of caution, numerical integration is a difficult computational problem especially if the integrand function is not unimodal. Although numerical integration may be a building procedure of commercial software we should use it carefully because typically no error analysis is supplied [4]. Also the reader should be aware that integral approximation may not guarantee the desired accuracy unless an analytical investigation of the integrand and its derivatives is carried out, especially when the integration domain is (-∞,∞). A good practice is to reestimate the density with different integration parameters to see whether estimation results are negligibly different. In other words, the sensitivity of estimation results to the choice of *Q*, z_q, and ω_q should be addressed in applications. Using approximation (3), we minimize the negative log likelihood function,

$$F(\alpha) = n \ln\left(\sum_{q=1}^{Q} w_q e^{-\Sigma_{k=1}^{K} \alpha_k P_k(z_q)} \right) + \sum_{i=1}^{n} \sum_{k=1}^{K} \alpha_k P_K(y_i). \tag{5}$$

Letting $p_q = w_q e^{-\Sigma_{k=1}^{K} \alpha_k P_k(z_q)}$, the derivatives are

$$\frac{\partial F}{\partial \alpha_j} = \sum_{i=1}^{n} P_j(y_i) - n \frac{\Sigma_{q=1}^{Q} P_j(z_q) p_q}{\Sigma_{q=1}^{Q} p_q}, \tag{6}$$

$$\frac{\partial^2 F}{\partial \alpha_j \partial \alpha_h} = \frac{n}{\left(\Sigma_{q=1}^{Q} p_q\right)^2} \times \left[\sum_{q=1}^{Q} P_k(z_q) P_h(z_q) p_q \sum_{q=1}^{Q} p_q - \sum_{q=1}^{Q} P_k(z_q) p_q \sum_{q=1}^{Q} P_h(z_q) p_q \right]. \tag{7}$$

The MLE solves equations $\partial F / \partial \alpha_j = 0$, $j = 1, ...,K$. The elements (7) constitute the $K \times K$ Hessian matrix H. We prove that this matrix is positive definite. Before that, we rewrite the gradient and the Hessian in matrix form. Let y be the $K \times 1$ vector with components $\Sigma_{i=1}^{n} P_j(y_i)$, Z be the Q ×K matrix with elements $Z_{qj} = P_j(z_q)$ and *p* be the $Q \times 1$ vector with components p_q. Then the gradient g, defined by (6) and the Hessian H, defined by (7), are written as

$$g = y - \frac{n}{p' 1_Q} Z' p,$$

$$H = \frac{n}{(p' 1_Q)^2} \left[Z' DZ \times (p' 1_Q) - Z' p p' Z \right],$$

where D = *diag*(p).

We prove that H is a positive-definite matrix. Let u be any nonzero $K \times 1$ vector, then u′Hu is proportional to v′Dv×(p′1_Q)-(v′p)^2, where v = Z′u. Rewriting this in coordinate form, we have

$$\left(\sum_{j=1}^{K} \upsilon_j^2 p_j \right) \left(\sum_{j=1}^{K} p_j \right) - \left(\sum_{j=1}^{K} \upsilon_j p_j \right)^2 \geq 0,$$

as follows from the Cauchy inequality expressing $v_j p_j = v_j \sqrt{p_j} \sqrt{p_j}$. The equality holds if and only if $v_j \sqrt{p_j} = \lambda \sqrt{p_j}$, i.e., if $v_j = const$. But Z′u = *const* would mean that the polynomial of the Kth order has *Q* roots which is impossible. Thus, the Hessian H is positive definite if $K < Q$. Consequently, the maximum likelihood estimate is unique because (5) is a strictly convex function. The Hessian, H, is also the Fisher information matrix because H is not random. The inverse, H^{-1} is the asymptotic covariance matrix of $\hat{\alpha}_{ML}$, which is used for hypothesis testing. The MLE is found iteratively by the Fisher Scoring algorithm,

$$a_{s+1} = a_s - H_s^{-1} g_s \quad (8)$$

where $s = 0, 1, \ldots$ is the iteration index and the right-hand side is evaluated at $\alpha = a_s$. As the iterations go on, the gradient $\|g_s\|$ should vanish.

A good test for numerical quadrature is to run the algorithm with $K = 2$ and to compare the solution to the exact one. Indeed if $K = 2$, we obtain the normal distribution with the negative log-likelihood value $0.5n\left[\ln\left(2\pi\hat{\sigma}^2\right)+1\right]$, where $\hat{\sigma}^2 = n^{-1}\Sigma_{i=1}^{n}\left(y_i - \bar{y}\right)^2$. This test may serve as a guide for choosing the right number of nodes.

Initial estimate

To start iterations (8), one has to have an initial guess for the vector of coefficients, a_0. The choice of the initial estimate is important because the closer a0 is to the MLE the faster iterations. We suggest the following procedure: Build a histogram with a chosen number of bins, $L > K$ at locations $\{y_l, l = 1, 2, \ldots, L\}$ and fit -ln p_l with $\{P_k(y_l), k = 1, \ldots, K\}$ using linear least squares, namely, minimizing

$$\sum_{l=1}^{L}(-\ln p_l - \alpha_0 - \alpha_1 P_1(y_l) - \alpha_2 P_2(y_l) - \ldots - \alpha_K P_K(y_l))^2 \quad (9)$$

The least squares estimate yields the initial vector $a_0 = (\hat{\alpha}_1 \ldots, \hat{\alpha}_K)'$. Several tries with different numbers of bins may be required to obtain satisfactory estimates, meaning that the coefficient at the highest order polynomial is positive and statistically significant. This procedure may help in determining the right polynomial degree via statistical significance testing of its coefficients.

Examples

We illustrate our approach with two examples. The first example has a moderate sample size, $n = 270$, with evident bimodal distribution, while in the second example, the sample size is fairly large, $n = 1057$, with a seemingly lognormal distribution. We use the statistical hypothesis approach to test normality and bimodality.

Rat brain oxygen distribution

We use the data from [10] on the rat brain PtO_2 measured with an Eppendorf polarographic electrode device. Knowing the distribution of oxygen in brain is a fundamental biological problem especially in connection with ischemia and lack of oxygen during stroke [7]. The histogram with 25 bins and three methods of density estimation using an exponential polynomial of the sixth order and the Gaussian kernel density is presented in Figure 1. We use the S-Plus built-in function density with the default bandwidth for the density estimation with Gaussian kernel. The bimodality of the brain oxygen distribution is visually obvious.All methods produce somewhat close density curves. The results of estimation are presented in Table 1 (SE are shown below in parentheses). *Initial* estimation uses linear fit of the sixth degree polynomial to the log frequency values by least squares (9). The nodes in *y-value* method are the observations themselves and G-H (21) refers to Gauss-Hermite numerical quadrature with 21 nodes. We test the quality of numerical quadrature by running the algorithms with K = 2. For the y-value method, the minimum F value (5) is 1019.52, while for the Gauss-Hermite numerical quadrature with 21 nodes, the value coincides with the exact one, 1033.53. Thus, we infer that the Gauss-Hermite numerical quadrature with 21 nodes yields a precise integral approximation.

Now we describe the estimation of the standard errors of mode and stable points. Let the vector of coefficients of the polynomial

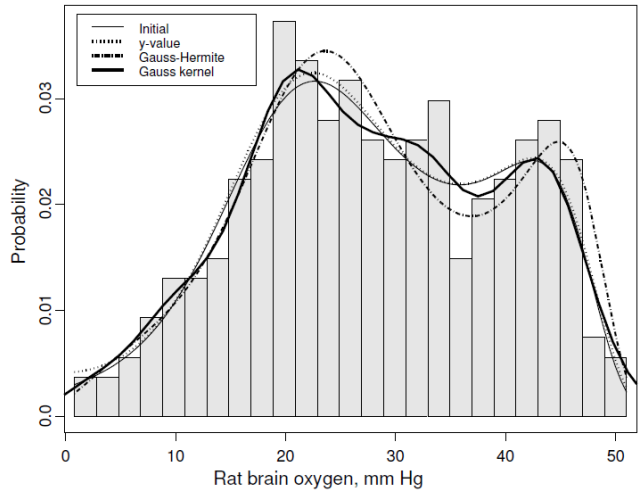

Figure 1: Histogram with 25 bins and four methods for density estimation.

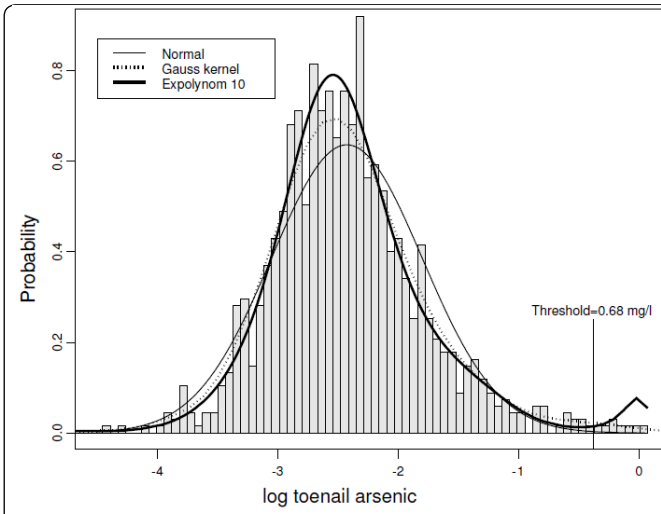

Figure 2: Distribution of toenail srsenic on the log scale. The distribution of New Hampshire residents with elevated arsenic starts at 0.68 mg/1.

of the K^{th} order be $\hat{\beta}$ with covariance matrix C = H^{-1}. Since modes and stable points are the roots of the polynomial $\Sigma_{k=1}^{K}\hat{\beta}_k k y^{k-1}$ we can estimate the variance of the root y_* with the delta method, $\left(\partial y_*/\partial\hat{\beta}_k\right)' C\left(\partial y_*/\partial\hat{\beta}_k\right)$, where the derivative of the root with respect to polynomial coefficients is calculated through the derivation of the implicit function, namely

$$\frac{\partial y_*}{\partial\hat{\beta}_k} = -\frac{\partial(\Sigma_{k=1}^{K}\hat{\beta}_k k y^{k-1})/\partial\hat{\beta}_k}{\partial(\Sigma_{k=1}^{K}\hat{\beta}_k k y^{k-1})/\partial_y} = -\frac{k y_*^{k-1}}{\Sigma_{k=2}^{K}\hat{\beta}_k k(k-1) y_*^{k-2}}. \quad (10)$$

The bimodality discovered has an important biological interpretation as an oxygen concentration in capillaries/blood vessels and brain matter. The saddle point may be used to discriminate PtO_2 values of highly and normally oxygenated parts of the brain. As follows from our density estimation, the oxygen oncentration in brain matter is half that of the blood vessels with the dividing value (saddle point value) at about 37 mm Hg.

Although bimodality and consequently abnormality is visually obvious, we need a statistical support by hypothesis testing. Testing normality is equivalent to testing the null hypothesis $H_0: \alpha_3 = \alpha_4 =$

Method	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	m#1	m#2	s.p.
Initial (25)	-1.256	2.845	-0.151	0.489	0.893	0.290	22.7	42.4	35.8
	(0.217)	(0.217)	(0.217)	(0.217)	(0.217)	(0.217)	(1.23)	(1.34)	(1.61)
y-value	-1.315	6.017	-1.165	1.192	2.973	0.706	20.6	43.6	35.4
	(1.002)	(1.004)	(1.002)	(1.090)	(1.152)	(1.072)	(1.74)	(1.37)	(1.83)
G-H (21)	-1.930	5.801	-2.302	0.587	2.698	1.902	22.1	45.3	36.7
	(1.005)	(1.023)	(1.061)	(1.025)	(0.817)	(0.671)	(1.16)	(0.79)	(0.92)

Note: m#1 – the left density peak/mode, m#2 = the right peak/mode, s.p. = saddle point

Table 1: Estimation results for rat brain PtO_2 density estimation.

$\alpha_5 = \alpha_6 = 0$. We test this hypothesis using the Wald and likelihood ratio test (ML Gauss-Hermite quadrature with 21 nodes is used in the following computations). Let C_* denote a 4×4 covariance matrix of $\hat{\alpha}_* = (\hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6)'$ as a sub matrix of the inverse to the Hessian, H^{-1}. Then, using the Wald test, we have $\hat{\alpha}_*' C_*^{-1} \hat{\alpha}_* \sim \chi^2(4)$, which gives the value 26 with the *p*-value less than 0.0001. For the likelihood ratio test, we compute $2(F_{2,\min} - F_{6,\min})$, where $F_{2,\min}$ and $F_{6,\min}$ are minimal values of (5) with $K = 2$and $K = 6$ respectively, which has $\chi^2(4)$ distribution. This test gives $2(F_{2,\min} - F_{6,\min}) = 37$with the *p*-value less that 0.0001. Thus, both tests confirm that the distribution of brain oxygen is not normal.

The advantage of the polynomial density estimation is obvious because it allows statistical hypothesis testing while the traditional kernel density estimation is used merely as the exploratory analysis. The former method of density estimation enables identifying the presence of two parts of the brain with statistically different levels of oxygen concentration.

Bump hunting: arsenic toenail distribution

Arsenic belongs to a group of toxic metals and may cause cancer. In particular, several papers relate the elevated concentration of arsenic in drinking water to bladder cancer by measuring the toenail arsenic concentration. Knowing the distribution of toenail arsenic is very important. For example, this distribution may help policy making to determine the threshold in the level of concentration above which the exposure to arsenic becomes dangerous for health. We hunt bumps at the right tail distribution [6].

Here we use data on toenail arsenic concentration in New Hampshire residents described in [8]. The histogram of the toenail arsenic distribution based on $n = 1057$ residents is 9 presented in Figure 2. The distribution is fairly close to lognormal. We show three density curves: normal, a nonparametric estimate with Gaussian kernel, and an exponential polynomial of the 10th degree. The latter density identifies a bump at right. Solving the polynomial equation of the ninth order, we find that the saddle point/threshold is -0.38. Using the delta-method with derivatives computed by formulas (10), we estimate the standard error of the threshold as 0.032.

Summing up, a group of New Hampshire residents have an elevated arsenic toenail concentration, which starts from 0.68 ml/l. Approximately 0.76% of population have an arsenic concentration that is dangerous for health. Having that threshold one may identify the area of the state where the concentration of arsenic is above the limit.

Discussion

Modeling the log density via orthogonal polynomials can be viewed as an alternative to the gold standard kernel density estimation. The choice of the kernel, such as triangle or Gaussian, is less important but the choice of the bandwidth is critical: If the bandwidth is wide the density looks like normal if the bandwidth is narrow the density is jagged. Although some methods of the bandwidth selection, like cross-validation, have been suggested in the literature they are ad-hoc in nature, and merely used for exploratory distribution analysis. The adventure of the polynomial density estimation is that it is model-based and therefore the classic statistical hypothesis testing applies. Consequently, we may rigorously answer the questions whether the distribution is normal or whether the discovered density bump is statistically significant. The log density estimation with orthogonal polynomials can be easily realized in modern statistical packages, such as S-Plus or R.

References

1. Brunk HD (1978) Univariate density estimation by orthogonal series. Biometrika 65: 521-528.
2. Buckland ST (1992) Fitting density functions with polynomials. Appl Stat 41: 63-76.
3. Clutton-Brock M (1990) Density estimation using exponential of orthogonal series. J Am Stat Assoc 85: 760-764.
4. Demidenko E (2004) Mixed Models: Theory and Applications. Jhon Wiley& Sons, New York.
5. Evans M, Swartz T (2000) Approximating Integrals via Monte Carlo and Deterministic Methods. Oxford University Press, Oxford.
6. Good IJ, Gaskins RA (1980) Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. J Am Stat Assoc 75: 42-56.
7. Hou H, Grinberg OY, Grinberg SA, Demidenko E, Swartz HM (2005) Cerebral tissue oxygenation in reversible focal ischemia in rats: Multi-site EPR oximetry measurements. Physiol Meas 26: 131-141.
8. Karagas MR, Tosteson TD, Morris JS, Demidenko E, Mott LA, et al. (2004) Incidence of transitional cell carcinoma of the bladder and arsenic exposure in New Hampshire. Cancer Causes Control 15: 465-472.
9. Kooperberg C, Stone CJ (1992) Logspline density estimation for censored data. J Comput Graph Stat 1: 301-328.
10. O'Hara JA, Khan N, Hou H, Wilmo CM, Demidenko E, et al. (2004) Comparison of EPR oximetry and Eppendorf polarographic electrode assessments of rat brain PtO2. Physiol Meas 25: 1413-1423.
11. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical Recipes in C. Cambridge University Press, New York.
12. Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.
13. Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, New York.