Journal of the Korean Data & Information Science Society
2009, **20**(5), 905–912

한국데이터정보과학회지

A kernel machine for estimation of mean and volatility functions[†]

Jooyong Shim[1] · Hyejung Park[2] · Changha Hwang[3]

[1]Department of Applied Statistics, Catholic University of Daegu
[2]Computer Course Division of General Education and Teacher's Certification, Daegu University
[3]Department of Statistics, Dankook University

Received 21 June 2009, revised 11 August 2009, accepted 20 August 2009

Abstract

We propose a doubly penalized kernel machine (DPKM) which uses heteroscedastic location-scale model as basic model and estimates both mean and volatility functions simultaneously by kernel machines. We also present the model selection method which employs the generalized approximate cross validation techniques for choosing the hyperparameters which affect the performance of DPKM. Artificial examples are provided to indicate the usefulness of DPKM for the mean and volatility functions estimation.

Keywords: Generalized approximate cross validation function, heteroscedastic regression, laplace distribution, location-scale model, penalized kernel regression.

1. Introduction

For given data set $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, with $\boldsymbol{x}_i \in \boldsymbol{R}^d$ and $y_i \in \boldsymbol{R}$, we consider the heteroscedastic regression model,

$$y_i = \mu(\boldsymbol{x}_i) + \sigma(\boldsymbol{x}_i)\epsilon_i \tag{1.1}$$

where $\boldsymbol{x}_i$ is the covariate vector and ϵ_i is assumed to follow a distribution with mean 0 and variance 1. The mean function $\mu(\boldsymbol{x}_i) = E(y_i|\boldsymbol{x}_i)$ and the volatility function $\sigma(\boldsymbol{x}_i) = \sigma(y_i|\boldsymbol{x}_i)$ (or the variance function $\sigma^2(\boldsymbol{x}_i)$) are to be estimated. Most nonparametric regression methods focus on the estimating the conditional mean for various data types.(Shim and Seok, 2008, Shim *et al.*, 2009). However estimating the variance function also is known important

[†] This research was supported by Korea SW Industry Promotion Agency (KIPA) under the program of Software Engineering Technologies Development and Experts Education.

[1] Adjunct Professor, Department of Applied Statistics, Catholic University of Daegu, Gyungbuk 712-702, Korea.
[2] Visiting Professor, Computer Course, Division of General Education and Teacher's Certification, Daegu University, Gyungbuk 712-714, Korea.
[3] Corresponding author: Full Professor, Department of Statistics, Dankook University, Gyeonggido 448-160, Korea. E-mail: chwang@dankook.ac.kr

in many studies (Gallant and Tauchen, 1997). The variance function is estimated based on the regression residuals previously obtained from differences of responses and estimates of the mean function. (Ruppert *et al.*, 1997; Fan and Yao, 1998). Bayesian approach has been introduced by Yau and Kohn (2003). Doubly penalized likelihood estimation based on the normal distribution has been proposed by Yuan and Wahba (2004). A distinguishing feature of it is estimating both the mean function and the variance function simultaneously without parametric assumption of either.

In this paper, we propose a doubly penalized kernel machine (DPKM) to take the heteroscedasticity into account and estimate both the mean function and the volatility function simultaneously. Laplace distribution is assumed to employ the robustness to the mean function estimation. The rest of this paper is organized as follows. The DPKM is introduced in Section 2, we propose a iteratively reweighted least squares (IRWLS) procedure for the mean function estimation and present Newton Raphson method for the volatility function estimation, and present the model selection method using the generalized approximate cross validation (GACV) functions. In Section 3 we perform the numerical studies through examples. In Section 4 we give the conclusions.

2. Doubly penalized kernel machine

We here consider the location-scale model based DPKM which estimates the mean function and the volatility function simultaneously. From the heteroscedastic regression model (1.1), we assume that ϵ_i follows independently double exponential distribution with mean 0 and scale parameter $1/\sqrt{2}$. $\mu(\boldsymbol{x}_i)$ and $\sigma(\boldsymbol{x}_i)$ are the mean function and the volatility function of y_i ,respectively, which are to be estimated. The negative log likelihood of the given data set can be expressed as (constant terms are omitted)

$$L(\mu, \sigma) = \frac{1}{n}\sum_{i=1}^{n} \frac{\sqrt{2}|y_i - \mu(\boldsymbol{x}_i)|}{\sigma(\boldsymbol{x}_i)} + \log\sigma(\boldsymbol{x}_i). \tag{2.1}$$

Due to the positivity of the volatility function we write the logarithm of $\sigma(\boldsymbol{x}_i)$ as $g(\boldsymbol{x}_i)$, then the negative log likelihood can reexpressed as

$$L(\mu, g) = \frac{1}{n}\sum_{i=1}^{n}\{\sqrt{2}|y_i - \mu(\boldsymbol{x}_i)|e^{-g(\boldsymbol{x}_i)} + g(\boldsymbol{x}_i)\}. \tag{2.2}$$

The mean function is estimated by a linear model, $\mu(\boldsymbol{x}) = \boldsymbol{\omega}_\mu{}'\boldsymbol{\phi}_\mu(\boldsymbol{x}) + b_\mu$ with a bias b_μ, conducted in a high dimensional feature space. Here the feature mapping function $\boldsymbol{\phi}_\mu(\cdot) : \boldsymbol{R}^d \to \boldsymbol{R}^{d_f}$ maps the input space to the higher dimensional feature space where the dimension d_f is defined in an implicit way. It is known that $\boldsymbol{\phi}_\mu(\boldsymbol{x}_i)'\boldsymbol{\phi}_\mu(\boldsymbol{x}_j) = \boldsymbol{K}_\mu(\boldsymbol{x}_i, \boldsymbol{x}_j)$ which are obtained from the application of Mercer (1909)'s conditions. Also g is estimated by a linear model, $g(\boldsymbol{x}) = \boldsymbol{\omega}_g{}'\boldsymbol{\phi}_g(\boldsymbol{x}) + b_g$ with a bias b_g.

Then the estimates of $(\mu, g, \boldsymbol{\omega}_\mu, \boldsymbol{\omega}_g)$ are obtained by minimizing the penalized negative log likelihood,

$$L(\boldsymbol{\omega}_\mu, \boldsymbol{\omega}_g, b_\mu, b_g) = \sum_{i=1}^{n}\{|y_i - \boldsymbol{\omega}_\mu{}'\boldsymbol{\phi}_\mu(\boldsymbol{x}_i) - b_\mu|e^{-g(\boldsymbol{x}_i)} + \boldsymbol{\omega}_g{}'\boldsymbol{\phi}_g(\boldsymbol{x}_i) + b_g\} \tag{2.3}$$

$$+\frac{\lambda_\mu}{2}\|\boldsymbol{\omega}_\mu\|^2+\frac{\lambda_g}{2}\|\boldsymbol{\omega}_g\|^2$$

where λ_μ and λ_g are nonnegative constants which control the trade-off between the goodness-of-fit on the data and $\|\boldsymbol{\omega}_\mu\|^2$ and $\|\boldsymbol{\omega}_g\|^2$. The representation theorem Kimeldorf and Wahba (1971) guarantees the minimizer of the penalized negative log likelihood to be $\mu(\boldsymbol{x}) = \boldsymbol{K}_\mu\boldsymbol{\alpha}_\mu + b_\mu$ and $g(\mathrm{x}) = \boldsymbol{K}_g\boldsymbol{\alpha}_g + b_g$, for some vectors $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_g$.

Now the problem (2.3) becomes obtaining $(\boldsymbol{\alpha}_\mu, \boldsymbol{\alpha}_g, b_\mu, b_g)$ to minimize

$$L(\boldsymbol{\alpha}_\mu, \boldsymbol{\alpha}_g, b_\mu, b_g) = \sqrt{2}|y - \boldsymbol{K}_\mu\boldsymbol{\alpha}_\mu - b_\mu|e^{-\boldsymbol{K}_g\boldsymbol{\alpha}_g - b_g} + \mathbf{1}'(\boldsymbol{K}_g\boldsymbol{\alpha}_g + b_g) \tag{2.4}$$

$$+\frac{\lambda_\mu}{2}\boldsymbol{\alpha}_\mu{}'\boldsymbol{K}_\mu\boldsymbol{\alpha}_\mu+\frac{\lambda_g}{2}\boldsymbol{\alpha}_g{}'\boldsymbol{K}_g\boldsymbol{\alpha}_g$$

where $\mathbf{1}$ is $n \times 1$ vector of 1's. The parameters $(\boldsymbol{\alpha}_\mu, \boldsymbol{\alpha}_g, b_\mu, b_g)$ of the conditional mean and volatility models can be found via an IRWLS procedure, alternating updates of the mean and volatility models.

2.1. Updating the conditional mean model

Fixing $g = \hat{g}$, the equation (2.4) reduces to

$$L(\boldsymbol{\alpha}_\mu, b_\mu) = \sqrt{2}|y - \boldsymbol{K}_\mu\boldsymbol{\alpha}_\mu - b_\mu|e^{-\hat{g}} + \frac{\lambda_\mu}{2}\boldsymbol{\alpha}_\mu{}'\boldsymbol{K}_\mu\boldsymbol{\alpha}_\mu. \tag{2.5}$$

The solution to (2.5) can be obtained by a weighted support vector machine (Vapnik, 1995, 1998) since (2.5) is actually equivalent to the objective function of a weighted support vector machine with weights$\sqrt{2}e^{-\hat{g}}$.

For easy selection of the optimal values of hyperparameters (λ_μ and other tuning parameters included in the kernel $\boldsymbol{K}_\mu$), we should not use the leave-one-out cross validation (LOO-CV) function but GACV function. We use IRWLS procedure so that the final estimator of $(\boldsymbol{\alpha}_\mu, b_\mu)$ can be expressed as the product of the hat matrix and y, which enables to obtain GACV function for the given values of hyperparameters.

We propose an IRWLS procedure to find the minimizers of (2.5) with a modified absolute loss function which is differentiable at 0. The modified absolute loss function $h_\delta(\cdot)$ is attained by providing the differentiability at 0 by differing from the original absolute loss function $h(\cdot)$ in the small interval $(-\delta, \delta)$,

$$h_\delta(r) = |r|I(|r| > \delta) + \frac{r^2}{\delta}I(|r| \leqq \delta) \tag{2.6}$$

where $\delta > 0$ and $I(\cdot)$ is an indicative function.

Now the likelihood function (2.5) becomes

$$L(\boldsymbol{\alpha}_\mu, b_\mu) = \sum_{i=1}^{n} u_i h_\delta(y_i - \boldsymbol{K}_{\mu i}\boldsymbol{\alpha}_\mu - b_\mu) + \frac{\lambda_\mu}{2}\boldsymbol{\alpha}_\mu{}'\boldsymbol{K}_\mu\boldsymbol{\alpha}_\mu \tag{2.7}$$

where $u_i = \sqrt{2}\exp(-g_i)$ and $\boldsymbol{K}_{\mu i}$ is the ith row of $\boldsymbol{K}_\mu$. Taking partial derivatives of (2.7) with regard to $\boldsymbol{\alpha}_\mu$ and b_μ leads to the optimal values of $\boldsymbol{\alpha}_\mu$ and b_μ are obtained from

$$\begin{pmatrix} \boldsymbol{\alpha}_\mu \\ b_\mu \end{pmatrix} = \begin{pmatrix} \boldsymbol{UHK}_\mu + \lambda_\mu \boldsymbol{I} & \boldsymbol{UH1} \\ \boldsymbol{1'UHK}_\mu & \boldsymbol{1'UH1} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{UHy} \\ \boldsymbol{1'UHy} \end{pmatrix}. \tag{2.8}$$

Here $\boldsymbol{U}$ is a digonal matrix of u_i's and $\boldsymbol{H}$ is a diagonal matrix with the ith diagonal element h_{ii} obtained from the derivative of the modified absolute loss function as

$$h_{ii} = \frac{1}{r_i} I(r_i > \delta) - \frac{1}{r_i} I(r_i < \delta) + \frac{2}{\delta} I(|r_i| \leqq \delta) \tag{2.9}$$

where $r_i = y_i - \boldsymbol{K}_{\mu i}\boldsymbol{\alpha}_\mu - b_\mu$.

The solution to (2.8) cannot be obtained in a single step since $\boldsymbol{H}$ contains $\boldsymbol{\alpha}_\mu$ and b_μ. Thus we need to apply IRWLS procedure which starts with initial values of $\boldsymbol{\alpha}_\mu$ and b_μ as follows:

(a) Calculate $\boldsymbol{H}$ with $\boldsymbol{\alpha}_\mu$ and b_μ .

(b) Calculate $\boldsymbol{\alpha}_\mu$ and b_μ from (2.8).

(c) Reiterate steps until convergence.

We now consider the cross validation (CV) function as follows:

$$CV(\theta) = \frac{1}{n}\sum_{i=1}^{n} u_i h_\delta(y_i - \hat{\mu}_\theta^{(-i)}(\boldsymbol{x}_i)) \tag{2.10}$$

where θ is the set of hyperparameters and $\hat{\mu}_\theta^{(-i)}(\boldsymbol{x}_i)$ is the estimate of $\mu(\boldsymbol{x}_i)$ estimated without ith observation. Since for each candidates of hyperparameters, $\hat{\mu}_\theta^{(-i)}(\boldsymbol{x}_i)$ for $i = 1, \cdots, n$, should be evaluated, selecting parameters using CV function is computationally formidable. By using a first order Taylor series expansion of the modified absolute loss function and the derivation procedure of GACV function from CV function by Yuan (2006), We have GACV function as follows

$$GACV(\theta) = \frac{\sum_{i=1}^{n} u_i h_\delta(y_i - \hat{\mu}_\theta^{(-i)}(\boldsymbol{x}_i))}{n - \mathrm{trace}(\boldsymbol{S})}, \tag{2.11}$$

where $S = (\boldsymbol{K}, \boldsymbol{1}) \begin{pmatrix} \boldsymbol{UHK} + \lambda_\mu \boldsymbol{I} & \boldsymbol{UHI} \\ \boldsymbol{1'UHK} & \boldsymbol{1'UH1} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{UH} \\ \boldsymbol{1'UH} \end{pmatrix}$ is the hat matrix such that $\hat{\boldsymbol{\mu}} = \boldsymbol{Sy}$.

2.2. Updating the conditional volatility model

Fixing $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$, the equation (2.4) reduces to

$$L(\boldsymbol{\alpha}_g, b_g) = \boldsymbol{z}' \exp(-\mathrm{K}_g\boldsymbol{\alpha}_g - b_g) + \boldsymbol{1}'(\boldsymbol{K}_g\boldsymbol{\alpha}_g + b_g) + \frac{\lambda_g}{2}\boldsymbol{\alpha}_g{}'\boldsymbol{K}_g\boldsymbol{\alpha}_g, \tag{2.12}$$

where $\boldsymbol{z}$ is $n \times 1$ vector with the ith elememt $\sqrt{2}|y_i - \hat{\mu}_i|$.

It is worth noting that (2.12) has the form of a penalized Gamma likelihood as if $z_i, i = 1, \cdots, n$ were independently drawn from Gamma distributions with shape parameter 1 and scale parameter $\exp(g_i) = \exp(\boldsymbol{K}_{gi}\boldsymbol{\alpha}_g + b_g), i = 1, \cdots, n$. This connection makes it possible to apply the general methodology for solving penalized likelihood problems with responses from exponential family. The model of the conditional standard deviation can then be updated efficiently via a simple Newton-Raphson method. In this step, the hyperparameters are λ_g and other tuning parameters included in the kernel $\boldsymbol{K}_g$. The hyperparameters are selected by GACV technique developed by Xiang and Wahba (1996)

$$GACV(\theta) = \frac{1}{n}\sum_{i=1}^{n}(z_i e^{-g_i} + g_i) + \frac{tr(\boldsymbol{V})}{n}\frac{\sum_{i=1}^{n} z_i(z_i - \exp(g_i))}{n - tr(\boldsymbol{W}^{1/2}\boldsymbol{V}\boldsymbol{W}^{1/2})}, \tag{2.13}$$

where $\boldsymbol{W}$ is a diagonal matrix whose ith element is $\exp(2g_i)$ and $\boldsymbol{V} = (\boldsymbol{W} + \boldsymbol{K}_g/\lambda_g)^{-1}$.

Summing up, we describe the algorithm for training and model selection of the DPKM for the heteroscedastic regression model as follows:

1) With given values of $\hat{g} = \boldsymbol{K}_g\hat{\boldsymbol{\alpha}}_g + \hat{b}_g$,

2) By GACV funtion (2.13), find the optimal values of λ_μ and other tuning parameters included in the kernel $\boldsymbol{K}_\mu$.

3) Find $\hat{\boldsymbol{\alpha}}_\mu$ and $\hat{b}_\mu$ from (2.8).

4) With $\hat{\boldsymbol{\mu}} = \boldsymbol{K}_\mu\hat{\boldsymbol{\alpha}}_\mu + \hat{b}_\mu$, by GACV funtion (2.13), find the optimal values of λ_g and other tuning parameters included in the kernel $\boldsymbol{K}_g$.

5) Find $\hat{\boldsymbol{\alpha}}_g$ and $\hat{b}_g$ from (2.12) using Newton-Raphson method.

6) Iterate 1-5 until convergence.

3. Numerical studies

We illustrate the performance of DPKM based on Laplace distribution using IRWLS procedure through two simulated data sets.

Example 1: For the first simulated example, $150(x_i, y_i)$'s are generated to present the estimation performance of the proposed method such that

$$y_i = f(x_i) + e_i = 2 + \sin(2\pi x_i) + e_i, i = 1, \cdots, 150,$$

where x_i is generated from the uniform distribution(0,1) and e_i is generated from Laplace distribution with mean 0 and scale parameter $\frac{1}{\sqrt{2}}\exp(x_i)$ (volatility of y_i is $\exp(x_i)$). The Gaussian kernel function and $\delta = 0.000001$ are utilized for the mean function estimation and the linear kernel function are utilized for the mean function estimation and volatility function estimation. Figure 3.1 (Left) shows true mean (solid line) and estimated mean function (dashed line) imposed on the scatter plots of 150 data points of y_i's in a data set.

Figure 3.1 (Right) shows true volatility (solid line) and estimated volatility function (dashed line) of a data set. We generated 100 data sets and obtained the MSEs for the performance metric as follows,

$$MSE_\mu = \frac{1}{150}\sum_{i=1}^{150}(\hat{\mu}_i - f(x_i))^2 \quad and \quad MSE_\sigma = \frac{1}{150}\sum_{i=1}^{150}(\hat{\sigma}_i - \sigma_i)^2.$$

We obtained the average of 100 MSE_μ's and their standard error as 0.081 and 0.0048, respectively. Also the average of 100 MSE_μ's and their standard error were obtained as 0.0351 and 0.0043, respectively.

Example 2: For the second simulated example, $150(x_i, y_i)$'s are generated to present the estimation performance of the proposed method such that

$$y_i = f(x_i) + e_i = 2(\exp(-30(x_i - 0.25)^2 + \sin(2\pi x_i)) - 2 + e_i, i = 1, \cdots, 150,$$

where $x_i = (i - 0.5)/150$ and e_i is generated from Laplace distribution with mean 0 and scale parameter $\frac{1}{\sqrt{2}}\exp(0.5\sin(2\pi x_i))$ (volatility of y_i is $\exp(0.5\sin(2\pi x_i))$). The Gaussian kernel function and $\delta = 0.000001$ are utilized for the mean function estimation and the Gaussian kernel function are utilized for the mean function estimation and volatility function estimation. Figure 3.2 (Left) shows true mean (solid line) and estimated mean function (dashed line) imposed on the scatter plots of 150 data points of y_i's. Figure 3.2 (Right) shows true volatility (solid line) and estimated volatility function (dashed line) of a data set. We obtained the average of 100 MSE_μ's and their standard error as 0.0572 and 0.0038, respectively. Also the average of 100 MSE_σ's and their standard error were obtained as 0.0367 and 0.0026, respectively.

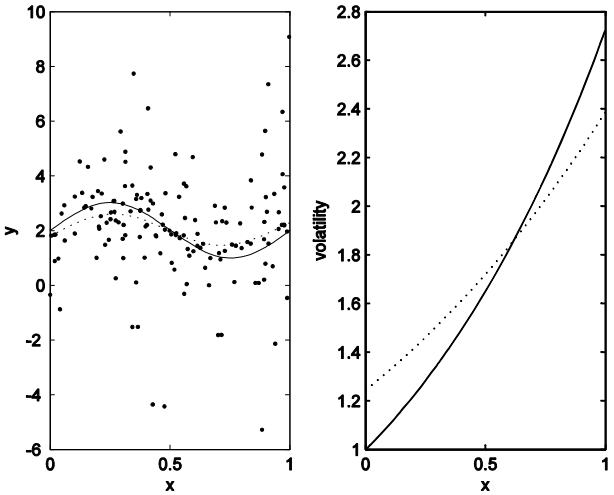

Figure 3.1 Mean function estimation (Left) and Volatility function estimation (Right) of a data set in Example 1.

From MSE's we can see that the proposed method provides the accurate estimation of mean and volatility functions, and from figures we can see that the estimated mean and volatility functions by proposed method behave similarly as the true functions do.

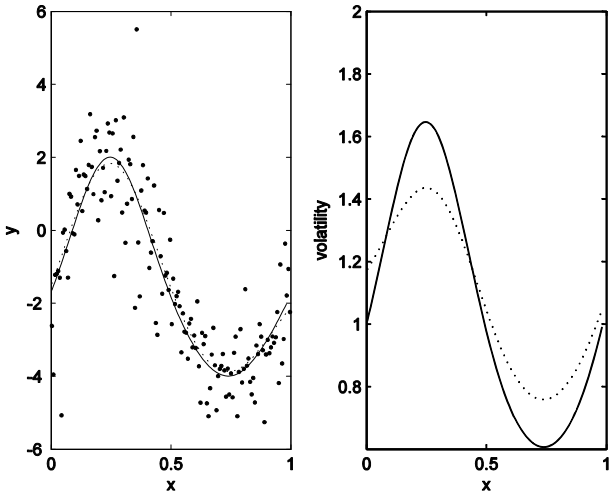

Figure 3.2 Mean function estimation (Left) and Volatility function estimation (Right) of a data set in Example 2.

4. Conclusions

In this paper, we dealt with estimating the mean function and the volatility function simultaneously by DPKM based on Laplace distribution. Through the examples we showed that the proposed procedure derives the satisfying results. We found that the doubly penalized kernel machine using IRWLS procedure provides the faster computation in training and model selection than using the weighted support vector machine for the mean function estimation.

References

Fan, J. Q. and Yao, Q. W. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**, 645-660.

Gallant, A. R. and Tauchen, G. (1997). Estimation of continuous time models for stock returns and interest rates. *Macroeconomic Dynamics*, **1**, 135-68.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and its Applications*, **33**, 82-95.

Mercer, J. (1909). Functions of positive and negative type and their connection with theory of integral equations. *Philosophical Transactions of Royal Society*, **A**, 415-446.

Ruppert, D., Wand, M. P., Holst, U. and Hossjer, O. (1997). Local polynomial variance-function estimation. *Technometrics*, **39**, 262-73.

Shim, J. and Seok, K. H. (2008). Kernel poisson regression for longitudinal data. *Journal of the Korean Data & Information Science Society*, **19**, 1353-1360.

Shim, J., Kim, T. Y., Lee, S. Y. and Hwang, C. (2009). Credibility estimation via Kernel mixed effects model. *Journal of Korean Data & Information Science Society*, **20**, 445-452.

Vapnik, V. N. (1995). *The nature of statistical learning theory*, Springer, New York.

Vapnik, V. N. (1998). *Statistical learning theory*, John Wiley, New York.

Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica*, **6**, 675-692.

Yau, P. and Kohn, R. (2003). Estimation and variable selection in nonparametric heteroscedastic regression. *Statistics and Computing*, **13**, 191-208.

Yuan, M. (2006). GACV for quantile smoothing splines. *Computational Statistics and Data Analysis*, **50**, 813-829.

Yuan, M. and Wahba, G. (2004). Doubly penalized likelihood estimator in heteroscedastic regression. *Statistics & Probability Letters*, **69**, 11-20.