UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK9030 — Modern Data Analysis
Day of examination:	Thursday December 13'th 2012
Examination hours:	14.30 – 18.30

This problem set consists of 2 pages.

Appendices:	None
Permitted aids:	Approved calculator

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1 (30% of total points)

Explain the main features of B-splines and describe how they are used.

Problem 2 (40% of total points)

Assume the model is linear of the form

$$Y = x^T\beta + \varepsilon$$

where $\mathbf{X}$ is the $N \times (p+1)$ matrix of inputs. There are therefore N observations, and the responses are collected in the N-dimensional vector $\mathbf{y}$. The error terms are independently Gaussian distributed with mean zero and variance σ_ε^2

The fitted values using ordinary least squares (OLS) are

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}.$$

a) Show that $trace(\mathbf{H}) = p + 1$.

b) Show that $\sum_{i=1}^N cov(\hat{y}_i, y_i) = (p+1)\sigma_\varepsilon^2$

c) Explain what is meant by a linear fitting method and the *effective degrees-of-freedom*.

d) Let $N_k(x)$ be the k closest points to the point x in some distance. Then

$$\hat{f}(x) = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i$$

is the k-nearest neighborhood estimator. Let now $\hat{f}(x_i),\ i = 1, \ldots, N$ be the fitted values. Show that $\sum_{i=1}^N cov(\hat{f}(x_i), y_i) = \frac{N}{k}$. Explain why nearest neighborhood estimation is a linear fitting method.

(Continued on page 2.)

Problem 3 (30% of total points)

Consider the situation where the response is categorical with K categories. The covariates or inputs are collected in a $N \times (p+1)$ matrix, $\mathbf{X}$.

a) Describe the logistic regression model in the case where the response is binary, i.e. $K = 2$, and derive an expression for the log-likelihood function when the distribution of the responses is binomial.

b) Explain how the model can be formulated for $K \geq 2$ categories, and derive the loglikelihood when the distribution of the responses is multinomial. Also indicate how a local likelihood can be formulated.

c) Show that fitting a locally constant multinomal logit model amounts to smooth the response indicators separately using the Nadaraya-Watson kernel smoother.

END