Maximum Likelihood Estimation

Consider a Gaussian process $X(s)$ observed at points $s_1, \ldots, s_n$, where $s \in \mathbb{R}^n$. The process is defined by a mean function $\mu(s) = d(s)'\beta$ and a covariance function $C(s, s')$. The inferential problem is that of maximizing the likelihood of the parameters that define μ and C, say β and ψ.

Maximum Likelihood Estimation

Consider a Gaussian process $X(s)$ observed at points $s_1, \ldots, s_n$, where $s \in \mathbb{R}^n$. The process is defined by a mean function $\mu(s) = d(s)'\beta$ and a covariance function $C(s, s')$. The inferential problem is that of maximizing the likelihood of the parameters that define μ and C, say β and ψ.

Given that the process is Gaussian we have that the vector $X = (X(s_1), \ldots, X(s_n))'$ is normally distributed with mean $D\beta$ and covariance matrix $V(\psi)$, where $V(\psi)_{ij} = C(s_i, s_j)$. Thus

$$L(\beta, \psi) \propto |V(\psi)|^{-1/2} \exp\left\{ -\frac{1}{2}(X - D\beta)'V(\psi)^{-1}(X - D\beta) \right\}$$

We can use sufficiency to write the likelihood as

$$L(\beta, \psi) \propto |V(\psi)|^{-1/2} \exp\left\{-\frac{1}{2}\left((\beta - \hat{\beta})' D' V(\psi)^{-1} D(\beta - \hat{\beta}) + S^2(\psi)\right)\right\}$$

where

$$D'V(\psi)^{-1} D\hat{\beta} = D'V(\psi)^{-1} X \quad \text{and} \quad S^2(\psi) = (X - D\hat{\beta})' V(\psi)^{-1} (X - D\hat{\beta})$$

This clearly shows that, for any given ψ the MLE of β is $\hat{\beta}$.

Maximum Likelihood Estimation

We can use sufficiency to write the likelihood as

$$L(\beta, \psi) \propto |V(\psi)|^{-1/2} \exp\left\{-\frac{1}{2}\left((\beta - \hat{\beta})' D' V(\psi)^{-1} D(\beta - \hat{\beta}) + S^2(\psi)\right)\right\}$$

where

$$D'V(\psi)^{-1}D\hat{\beta} = D'V(\psi)^{-1}X \quad \text{and} \quad S^2(\psi) = (X - D\hat{\beta})'V(\psi)^{-1}(X - D\hat{\beta})$$

This clearly shows that, for any given ψ the MLE of β is $\hat{\beta}$.

To calculate $\hat{\beta}$ we need a square root of V and a QR or SVD decomposition of D. Calculating V^{-1} and forming the product $D'V(\psi)^{-1}D$ is usually a VERY BAD idea.

$S^2(\psi, \beta)$ is obtained as a by-product of the LSE calculation for $\hat{\beta}$.

Consider a $n \times n$ orthogonal matrix P. Then $P^{-1} = P'$. Thus

$$x \in \mathbb{R}^n, \ \ ||Px||^2 = x'P'Px = x'x = ||x||^2$$

so P preserves the Euclidean norm of vectors. Intuitively, P is a rotation in $\mathbb{R}^n$.

Consider a $n \times n$ orthogonal matrix P. Then $P^{-1} = P'$. Thus

$$x \in \mathbb{R}^n, \;\; ||Px||^2 = x'P'Px = x'x = ||x||^2$$

so P preserves the Euclidean norm of vectors. Intuitively, P is a rotation in $\mathbb{R}^n$.

Consider the linear model $Y = D\beta + e$. Consider the QR decomposition of D given as $D = QR$, where Q is a $n \times b$ orthogonal matrix and R is $n \times k$ matrix with an upper triangular matrix R_1 in the upper $k \times k$ block and zeroes in the lower block.

Let $D = QR$. The LSE solution minimizes

$||Y - D\hat{\beta}||^2 = ||Y - QR\hat{\beta}||^2 = ||Q'Y - R\hat{\beta}||^2 = ||(Q'Y)_1 - R_1\hat{\beta}||^2 + ||(Q'Y)_2||^2$

This expression reaches a minimum when

$$R_1\hat{\beta} = (Q'Y)_1$$

This is a triangular system, so is very easy and fast to solve.

Let $D = QR$. The LSE solution minimizes

$$||Y - D\hat{\beta}||^2 = ||Y - QR\hat{\beta}||^2 = ||Q'Y - R\hat{\beta}||^2 = ||(Q'Y)_1 - R_1\hat{\beta}||^2 + ||(Q'Y)_2||^2$$

This expression reaches a minimum when

$$R_1\hat{\beta} = (Q'Y)_1$$

This is a triangular system, so is very easy and fast to solve.

Moreover, at the minimum

$$||Y - D\hat{\beta}||^2 = ||(Q'Y)_2||^2 \text{ and } D'D = R'Q'QR = R'R$$

so S^2 is equal to the norm of the last $n - k$ elements of the vector $Q'Y$ and R is the Cholesky factor of the covariance matrix $D'D$.

Suppose the regression error e is such that $\mathrm{var}(e) = V$. Then take the Cholesky decomposition $V = LL'$. Then

$$Y = D\beta + e \Rightarrow L^{-1}Y = L^{-1}D\beta + L^{-1}e \ \text{ or } \ Z = F\beta + \varepsilon$$

where $\mathrm{var}(\varepsilon) = L^{-1}VL^{-T} = L^{-1}LL'L^{-T} = I$.

Suppose the regression error e is such that $\mathrm{var}(e) = V$. Then take the Cholesky decomposition $V = LL'$. Then

$$Y = D\beta + e \Rightarrow L^{-1}Y = L^{-1}D\beta + L^{-1}e \ \text{ or } \ Z = F\beta + \varepsilon$$

where $\mathrm{var}(\varepsilon) = L^{-1}VL^{-T} = L^{-1}LL'L^{-T} = I$.

The normal equations for the transformed linear model are

$$F'F\hat{\beta} = F'Z \ \text{ or } \ D'L^{-t}L^{-1}F\hat{\beta} = D'L^{-T}L^{-1}Y$$

which is equal to $D'V^{-1}D\hat{\beta} = D'V^{-1}Y$.

Suppose the regression error e is such that $\mathrm{var}(e) = V$. Then take the Cholesky decomposition $V = LL'$. Then

$$Y = D\beta + e \Rightarrow L^{-1}Y = L^{-1}D\beta + L^{-1}e \;\text{ or }\; Z = F\beta + \varepsilon$$

where $\mathrm{var}(\varepsilon) = L^{-1}VL^{-T} = L^{-1}LL'L^{-T} = I$.

The normal equations for the transformed linear model are

$$F'F\hat{\beta} = F'Z \;\text{ or }\; D'L^{-t}L^{-1}F\hat{\beta} = D'L^{-T}L^{-1}Y$$

which is equal to $D'V^{-1}D\hat{\beta} = D'V^{-1}Y$.

So the computational method to obtain the solution to the weighed LSE consists of: (a) computing the Cholesky decomposition of the covariance matrix; (b) solving $LZ = Y$ and $LF = D$; (c) calculating the QR decomposition of F; (d) Solving for $\hat{\beta}$.

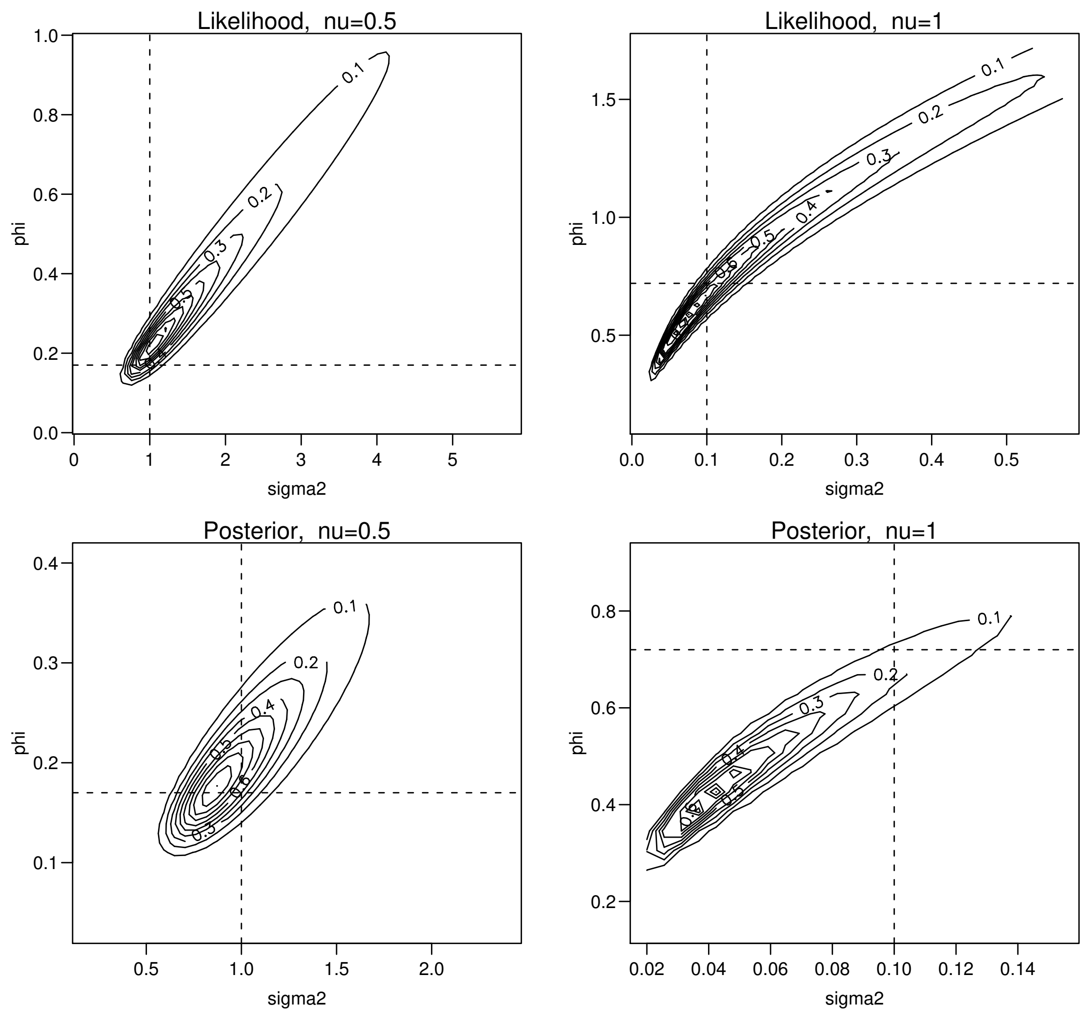
- Matrix computations will have to be performed repeatedly and have to be fast and accurate.
- Never invert a matrix explicitly.
- Prefer methods that are based on orthogonal transformations like QR or SVD.
- Use methods that account for the particular structure of the matrix. This is particularly important over large regular grids.
- Use simulation methods that can handle strong correlations between parameters.

Estimating Correlation Parameters

There are several problems related to the estimation of the correlation parameters using likelihood methods.

- Computational issues. Cholesky decompositions require $O(n^3)$ operations. This represents a heavy computational burden.
- Traditional MLE methods need the likelihood to be two times differentiable. If the covariance function does not satisfy this condition, then neither does the likelihood.
- The likelihood can be very flat, implying that different parameter values produce close to identical results.
- There can be strong correlations between the parameters that define the covariance function, resulting in "banana" shaped likelihood surfaces for which maximization is difficult.

Likelihood and posterior density functions for the range and the scale of two Matérn class correlations: $\nu = .5$ (exponential) and $\nu = 1$ (Whittle). Dotted lines correspond to true values.



To simplify the problem of visualizing the likelihood in multi-dimensional settings one can use **Profile Likelihoods**. Consider a likelihood depending on parameters (α, φ). Then

$$L_p(\alpha) = L(\alpha, \hat{\varphi}(\alpha)) = \max_{\varphi}(L(\alpha, \varphi))$$

To simplify the problem of visualizing the likelihood in multi-dimensional settings one can use **Profile Likelihoods**. Consider a likelihood depending on parameters (α, φ). Then

$$L_p(\alpha) = L(\alpha, \hat{\varphi}(\alpha)) = \max_{\varphi}(L(\alpha, \varphi))$$

An alternative to the maximum-based profile likelihood is the marginalized profile likelihood. So we look at $L_I(\alpha)$ which is obtained after integrating φ out. Integrating out σ^2 and β in a Gaussian process likelihood.

The integration proposed in the previous slide consists of

$$\int_{\mathbb{R}^k} \int_0^\infty d\beta d\sigma^2 |V(\psi)|^{-1/2} (\sigma^2)^{-n/2}$$

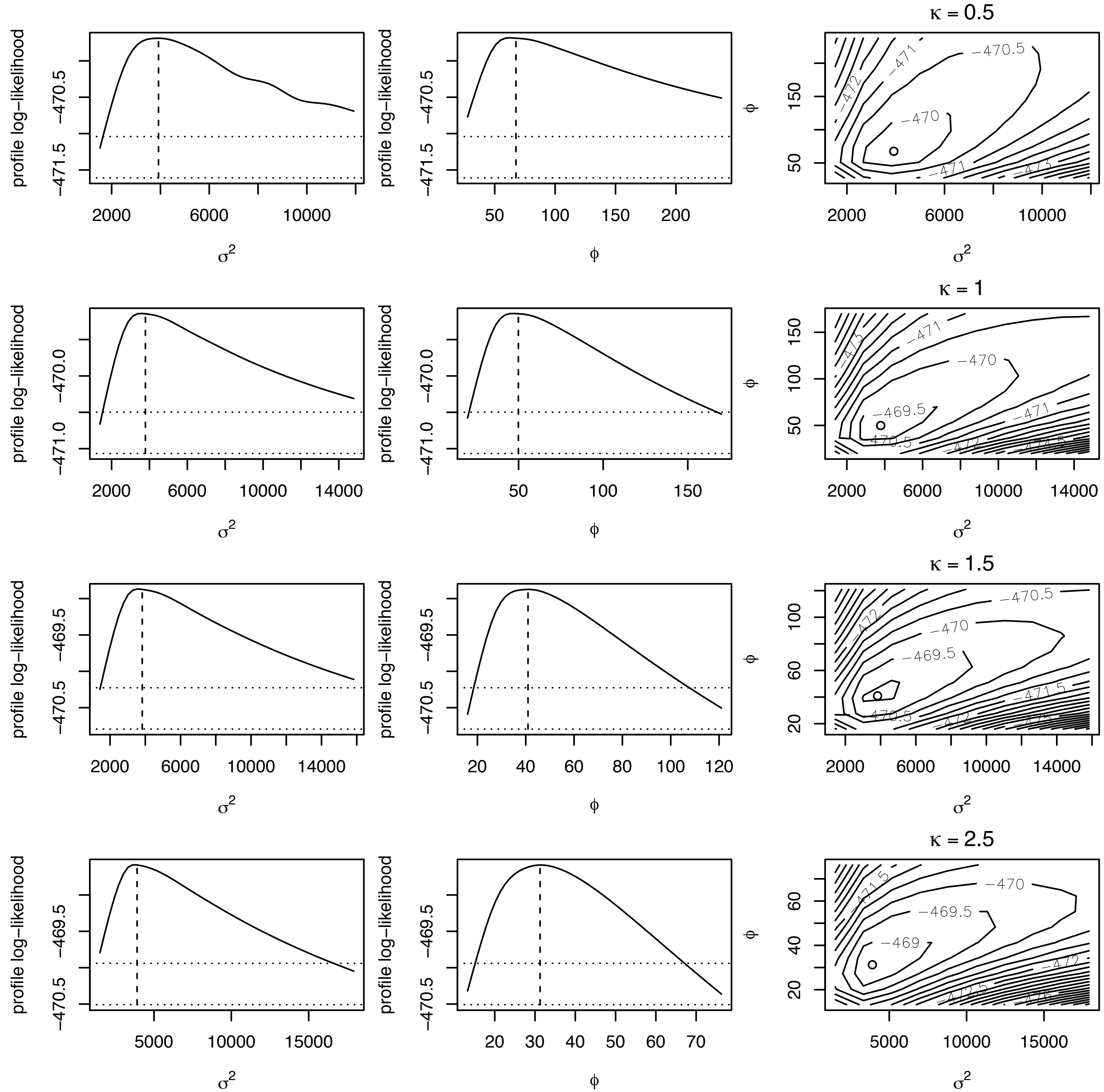
$$\exp\left\{-\frac{1}{2\sigma^2}\left((\beta - \hat{\beta})' D' V(\psi)^{-1} D(\beta - \hat{\beta}) + S^2(\psi)\right)\right\}$$

that yields

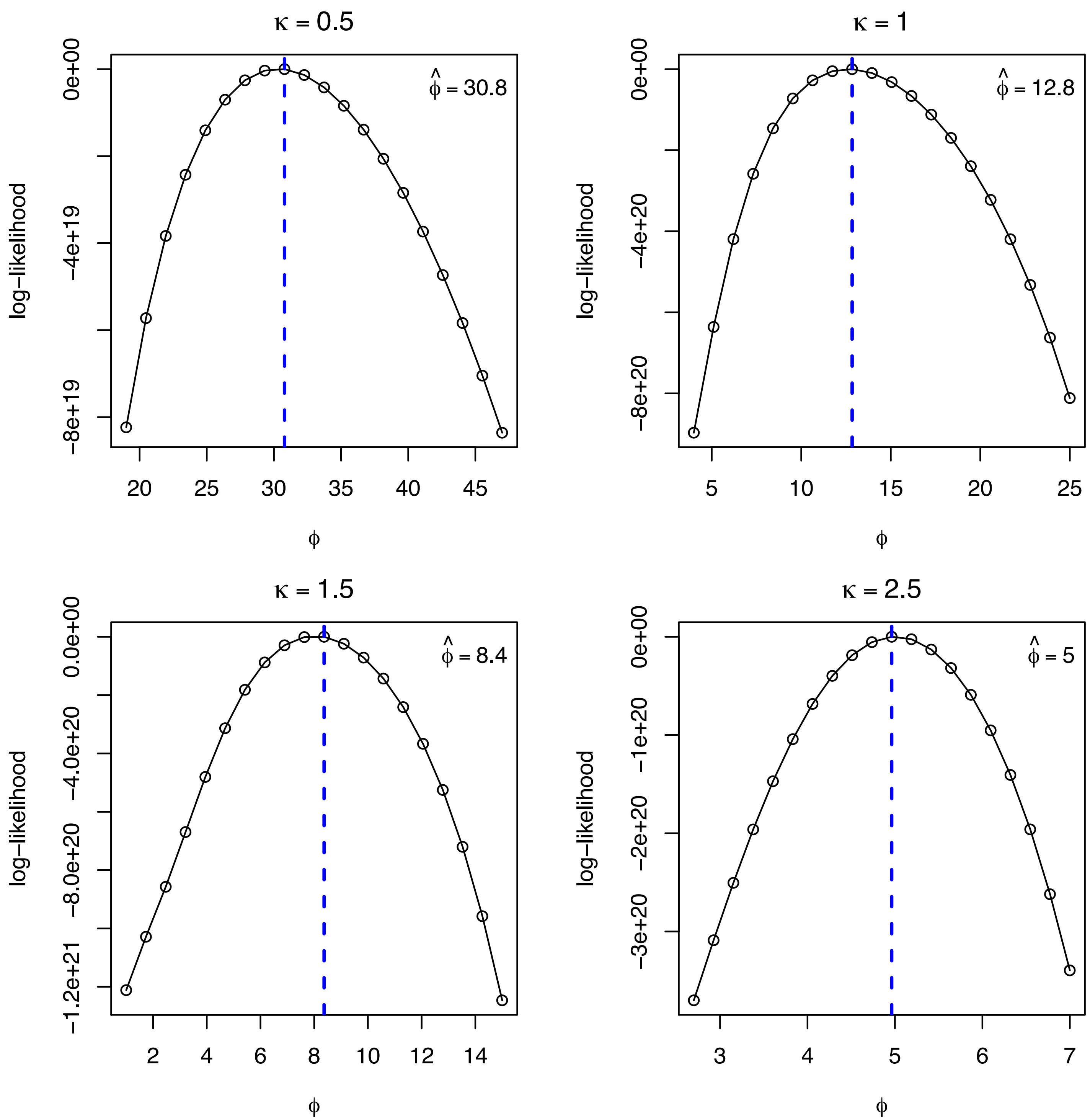
$$L(\psi) \propto |V(\psi)|^{-1/2} |D' V(\psi)^{-1} D|^{-1/2} (S(\psi)^2)^{-(m-k)/2}$$

where $\beta \in \mathbb{R}^k$ (Mardia and Watkins, Biometrika '89). This expression is much more regular than the original likelihood and can be used to obtain a maximum marginal likelihood estimator of ψ. It also plays a fundamental role in the non-subjective Bayes analysis of the problem.

Likelihood and profile likelihood for nugget = 6,000 and different values of the smoothness parameter in a Matèrn correlation family



Guárico Rainfall Marginal Likelihood



Marginal likelihood for the range parameter for sill to nugget ratio of 1.3 as suggested by the variogram, and different values of the smoothness parameter in a Matèrn correlation family

Consistency

Zhang, JASA 04 established that the sill and the range of a correlation function in the Matèrn family can not be estimated consistently.

The result is based on the fact that, for a given ν, two elements of the family produce equivalent probability measures if the ratio σ/ϕ^{ν} is the same for both of them.

Consistency

Zhang, JASA 04 established that the sill and the range of a correlation function in the Matèrn family can not be estimated consistently.

The result is based on the fact that, for a given ν, two elements of the family produce equivalent probability measures if the ratio σ/ϕ^ν is the same for both of them.

These results imply that only σ/ϕ^ν can be estimated consistently. A transformation that uses a function of σ/ϕ^ν can be used to improve the estimation, but it would not solve the consistency problem.