Statistica Sinica **23** (2013), 1131-1153
doi:http://dx.doi.org/10.5705/ss.2011.319

# NONPARAMETRIC DENSITY ESTIMATION IN HIGH-DIMENSIONS

Chong Gu, Yongho Jeon and Yi Lin

*Purdue University, Yonsei University and Verition Fund Management, LLC.*

*Abstract:* Penalized likelihood density estimation provides an effective approach to the nonparametric fitting of graphical models, with conditional independence structures characterized via selective term elimination in functional ANOVA decompositions of the log density. A bottleneck in the approach has been the cost of numerical integration, which has limited its application to low-dimensional problems. In Jeon and Lin (2006), a reformulation was proposed to replace multi-dimensional integrals by sums of products of univariate integrals, greatly reducing the numerical burden in high-dimensional problems. In this article, we derive a cross-validation score for use with the reformulation that delivers effective smoothing parameter selection at a manageable computational cost, introduce a geometric inference tool for the "testing" of model terms, and calculate the asymptotic convergence rates of the estimates. An assortment of practical issues are investigated through empirical studies, and open-source software is illustrated with data examples.

*Key words and phrases:* Cross-validation, graphical models, penalized likelihood, projection, smoothing parameter.

## 1. Introduction

Consider the nonparametric estimation of a probability density $p(x)$ on a domain $\mathcal{X}$ based on independent samples $X_i$, $i = 1, \ldots, n$. Numerous methods have been developed over the years, but most have found little practical success in dimensions beyond two or three. The immediate challenge in high dimensions is the curse of dimensionality, as multivariate functions are intrinsically difficult to estimate. One approach to easing the curse of dimensionality is via the elimination of higher order terms in certain ANOVA decompositions of multivariate functions.

On a product domain, say $\mathcal{X} = \prod_{j=1}^{d} \mathcal{X}_j$, one may decompose a function $\eta$ as

$$\eta(x) = \eta(x_1, \ldots, x_d) = \eta_\emptyset + \textstyle\sum_j \eta_j(x_j) + \sum_{j<k} \eta_{j,k}(x_j, x_k) + \ldots, \tag{1.1}$$

with the constant in $\eta_\emptyset$, the main effects in $\eta_j$, the two-way interactions in $\eta_{j,k}$, etc.; the function is easier to estimate when higher order interactions are not

involved. Such an ANOVA decomposition can be built into penalized likelihood density estimation that minimizes

$$-\frac{1}{n}\sum_{i=1}^{n}\eta(X_i)+\log\int_{\mathcal{X}}e^{\eta(x)}dx+\frac{\lambda}{2}J(\eta), \tag{1.2}$$

where $p(x)=e^{\eta(x)}/\int_{\mathcal{X}}e^{\eta(x)}dx$, $J(\eta)$ is a quadratic roughness functional, and the smoothing parameter $\lambda$ controls the tradeoff between the smoothness of $\eta(x)$ and its fidelity to the data.

The formulation of (1.2) can be found in Gu and Qiu (1993), which evolved from the original proposal of Good and Gaskins (1971) through works by Leonard (1978), Silverman (1982), and O'Sullivan (1988). The computation of (1.2) with cross-validated $\lambda$ has been studied in Gu (1993) and Gu and Wang (2003), where integrals of form $\int_{\mathcal{X}}h(x)e^{\eta(x)}dx$ have to be calculated while $\eta(x)$ is being updated iteratively. Numerical integration can be computationally prohibitive on high dimensional domains, however, limiting the practical applicability of the method.

To circumvent the computational hurdle associated with (1.2) while maintaining the versatility of the modular ANOVA structure in log density, Jeon and Lin (2006) proposed to minimize

$$\frac{1}{n}\sum_{i=1}^{n}e^{-\eta(X_i)}+\int_{\mathcal{X}}\eta(x)\rho(x)dx+\frac{\lambda}{2}J(\eta) \tag{1.3}$$

for $\rho(x)$ some known density satisfying $\int_{\mathcal{X}}\rho(x)dx=1$, with the resulting density estimate $\hat{p}(x)\propto e^{\hat{\eta}(x)}\rho(x)$. With proper choices of $\rho(x)$, the integral $\int_{\mathcal{X}}\eta(x)\rho(x)dx$ appearing in (1.3) can be decomposed into sums of products of univariate integrals, permitting fast computation.

For (1.3) to work in practice, a critical aspect is the automatic selection of $\lambda$ that delivers reasonable performance. A simple five-fold cross-validation was used for the purpose in the examples of Jeon and Lin (2006), which required the evaluation of the normalizing constant $\int_{\mathcal{X}}e^{\eta(x)}\rho(x)dx$, the very operation (1.3) was designed to avoid. The primary objective here is to develop a cross-validation scheme for use with (1.3) that does not involve $\int_{\mathcal{X}}e^{\eta(x)}\rho(x)dx$ or the like. Also presented is an asymptotic theory concerning the convergence rates of the minimizers of (1.3) that contributes to the understanding of the method and also helps its implementation.

Selective elimination of interaction terms in ANOVA decompositions of log density $\eta(x)$ may imply conditional independence structures, and estimation via (1.2) or (1.3) provides a means to the nonparametric fitting of graphical models. For example, on $\mathcal{X}_1\times\mathcal{X}_2\times\mathcal{X}_3$ for $p(x)=e^{\eta(x)}/\int_{\mathcal{X}}e^{\eta(x)}dx$, $\eta(x)=\eta_\emptyset+\eta_1+$

$\eta_2+\eta_3+\eta_{1,2}+\eta_{1,3}$ implies the conditional independence of $X_2$ and $X_3$ given $X_1$, or $X_2 \perp X_3|X_1$. To infer such ANOVA structures, Kullback-Leibler projection was proposed in Gu (2004) for the "testing" of insignificant interactions in log density, which however involved integrals of the form $\int_{\mathcal{X}} h(x)e^{\eta(x)}dx$. A certain squared error projection is also implemented here to accomplish the task without the offending numerical operation.

The rest of the article is organized as follows. In Section 2, technical details are filled in concerning the formulation and computation of (1.3). Smoothing parameter selection is discussed in Section 3, where a cross-validation score is derived to work in tandem with (1.3). Squared error projection is introduced in Section 4 for the "testing" of ANOVA terms. Simulation studies are conducted in Section 5 to assess an assortment of practical issues. In Section 6, data examples are presented to showcase potential applications of the techniques using open-source software. Asymptotic convergence is studied in Section 7, followed by miscellaneous remarks in Section 8.

## 2. Preliminaries

We fill in some specifics in the formulation, discuss basic properties, and set up the notation. Some of these were treated in greater details in Jeon and Lin (2006)

### 2.1. Reproducing kernel Hilbert spaces

The minimization of (1.3) is implicitly in a Hilbert space $\mathcal{H} \subseteq \{f : J(f) < \infty\}$ in which $J(f)$ is a square semi-norm with a finite dimensional null space $\mathcal{N}_J = \{f : J(f) = 0\}$. A Hilbert space has a metric and a geometry that facilitate analysis and computation, and a finite dimensional $\mathcal{N}_J$ prevents interpolation. Function evaluations appear in (1.3), so one also needs the evaluation functional $[x]f = f(x)$ to be continuous in $f \in \mathcal{H}$, $\forall x \in \mathcal{X}$.

A Hilbert space in which evaluation functional is continuous is a reproducing kernel Hilbert space with a reproducing kernel $R(\cdot, \cdot)$, a non-negative definite bivariate function on $\mathcal{X}$ such that $R(x, \cdot) = R(\cdot, x) \in \mathcal{H}$, $\forall x \in \mathcal{X}$, and $\langle R(x, \cdot), f(\cdot)\rangle = f(x)$ (the reproducing property), $\forall f \in \mathcal{H}$, where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{H}$. A reproducing kernel Hilbert space can be generated from its reproducing kernel $R$, for which any non-negative definite function qualifies, as the "column space" span$\{R(x, \cdot), x \in \mathcal{X}\}$. A general theory can be found in Aronszajn (1950).

In the settings of (1.2) and (1.3), one can write $\langle \cdot, \cdot \rangle = J(\cdot, \cdot) + \tilde{J}(\cdot, \cdot)$, where $J(\cdot, \cdot)$ is the semi inner product associated with $J(f)$ and $\tilde{J}(\cdot, \cdot)$ is an inner product in $\mathcal{N}_J$. One has a tensor-sum decomposition $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$, with $J(f)$ being a square full norm in $\mathcal{H}_J$. For computation, one needs a basis of $\mathcal{N}_J$ and

the reproducing kernel $R_J$ in $\mathcal{H}_J$ satisfying $J(R_J(x,\cdot), f(\cdot)) = f(x)$, $\forall f \in \mathcal{H}_J$, $\forall x \in \mathcal{X}$.

For an example, consider the cubic smoothing spline on $\mathcal{X} = [0,1]$ with $J(f) = \int_0^1 \big(f''(x)\big)^2 dx$. Taking $\tilde{J}(f,f) = \big(\int_0^1 f(x)dx\big)^2 + \big(\int_0^1 f'(x)dx\big)^2$, the reproducing kernel in $\mathcal{H}_J = \{f : J(f) < \infty, \int_0^1 f(x)dx = \int_0^1 f'(x)dx = 0\}$ is given by $R_J(x_1,x_2) = k_2(x_1)k_2(x_2) - k_4(|x_1 - x_2|)$, where $k_\nu = B_\nu/\nu!$ are scaled Bernoulli polynomials. Combining with $\mathcal{N}_J = \{1\} \oplus \big\{k_1(x)\big\}$, where $k_1(x) = x - 0.5$, one has a one-way ANOVA decomposition $\eta = \eta_\emptyset + \eta_x$, with $\eta_x \in \big\{k_1(x)\big\} \oplus \mathcal{H}_J$ satisfying the side condition $\int_0^1 \eta_x(x)dx = 0$. The construction provides building blocks for tensor-product cubic splines, to be discussed below, for which one writes $\mathcal{H} = \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_1$ with reproducing kernels $R_{00}(x_1,x_2) = 1$, $R_{01}(x_1,x_2) = k_1(x_1)k_1(x_2)$, and $R_1(x_1,x_2) = k_2(x_1)k_2(x_2) - k_4(|x_1 - x_2|)$.

On $\mathcal{X} = \mathcal{U} \times \mathcal{V} = [0,1]^2$, one may construct tensor-product cubic splines using the marginal construction given above, with nine tensor-sum terms $\mathcal{H}_{\mu,\nu} = \mathcal{H}_\mu^{(u)} \otimes \mathcal{H}_\nu^{(v)}$ on $\mathcal{X}$, $\mu,\nu = 00, 01, 1$, generated from reproducing kernels $R_{\mu,\nu}(x_1,x_2) = R_\mu(u_1,u_2)R_\nu(v_1,v_2)$. The four subspaces with $\mu,\nu = 00, 01$ are of one-dimension each, and can be lumped together as $\mathcal{N}_J$. The other five subspaces can be put together as $\mathcal{H}_J$ with the reproducing kernel

$$R_J = \theta_{00,1}R_{00,1} + \theta_{1,00}R_{1,00} + \theta_{01,1}R_{01,1} + \theta_{1,01}R_{1,01} + \theta_{1,1}R_{1,1}, \tag{2.1}$$

where $\theta_{\mu,\nu}$ are a set of extra smoothing parameters adjusting the relative weights of the roughness of different components; the roughness penalty associated with $R_J$ in (2.1) is of the form $J(f) = \sum_{\mu,\nu} \theta_{\mu,\nu}^{-1} J_{\mu,\nu}(f_{\mu.\nu})$, where $J_\beta(f_\beta)$ is associated with $R_\beta$ for $f_\beta \in \mathcal{H}_\beta$. The nine subspaces readily define the ANOVA decomposition of (1.1), with $\eta_\emptyset \in \mathcal{H}_{00,00}$, $\eta_u \in \mathcal{H}_{01,00} \oplus \mathcal{H}_{1,00}$, $\eta_v \in \mathcal{H}_{00,01} \oplus \mathcal{H}_{00,1}$, and $\eta_{u,v} \in \mathcal{H}_{01,01} \oplus \mathcal{H}_{1,01} \oplus \mathcal{H}_{01,1} \oplus \mathcal{H}_{1,1}$. To obtain an additive model, one removes $\mathcal{H}_{01,01}$ from $\mathcal{N}_J$ and sets $\theta_{1,01} = \theta_{01,1} = \theta_{1,1} = 0$. Note that for any given $\tilde{x} \in \mathcal{X}$, $\xi(x) = R_J(\tilde{x}, x)$ is a linear combination of functions of form $f(u)g(v)$; the same holds for functions in $\mathcal{N}_J$. Constructions in higher dimensions can be done recursively.

For a one-to-one mapping $p(x) \leftrightarrow \eta(x)$ in (1.2), one can set $\eta_\emptyset = 0$.

### 2.2. Choice of $\rho(x)$

The motivation for (1.3) is to avoid numerical integrations in high dimensions, and one aims to factorize the integral $\int_{\mathcal{X}} \eta(x)\rho(x)dx$ into sums of products of univariate integrals. It is thus necessary for $\rho(x)$ to factorize into a product of univariate densities; as an added benefit, conditional independence structures can be characterized via selective elimination of ANOVA terms in $\eta(x)$, just as with (1.2).

Obvious choices for the factors of $\rho(x)$ are marginal density estimates, for which one can use parametric fits such as the beta distributions used in Jeon and Lin (2006), or nonparametric fits such as the minimizers of (1.2) with cross-validated $\lambda$; the latter is used in our implementation.

### 2.3. Existence

Let $\{\phi_\nu\}_{\nu=1}^m$ be a basis of $\mathcal{N}_J$ and $S$ be the $n \times m$ matrix with the $(i,\nu)$th entry $\phi_\nu(X_i)$. Define $L(f) = n^{-1}\sum_{i=1}^n e^{-f(X_i)} + \int_{\mathcal{X}} f(x)\rho(x)dx$.

**Lemma 1.** *If $S$ is of full column rank, then $L(f)$ is strictly convex in $\mathcal{N}_J$ and $L(f) + \lambda J(f)$ is strictly convex in $\mathcal{H}$.*

**Proof.** For $f, g \in \mathcal{H}$ and $\alpha \in (0,1)$, define $A(\alpha) = L(g + \alpha(f-g))$. It is easy to verify that $d^2A/d\alpha^2 = n^{-1}\sum_{i=1}^n e^{-(g+\alpha(f-g))(X_i)}(f-g)^2(X_i) \geq 0$, where the equality holds if and only if $(f-g)(X_i) = 0$, $i = 1,\ldots,n$; for $f - g \in \mathcal{N}_J$, the equality implies $f = g$ as $S$ is of full column rank. This, combined with the strict convexity of $J(f)$ in $\mathcal{H}_J$, establishes the lemma.

**Theorem 1.** *Suppose $S$ is of full column rank. If $L(f)$ has a unique minimizer in $\mathcal{N}_J$, then $L(f) + \lambda J(f)$ has a unique minimizer in $\mathcal{H}$.*

The theorem follows Lemma 1 and Theorem 2.9 in Gu (2002).

### 2.4. Equivalent formulation and computation

Suppose $J(f)$ annihilates constant and consider a tensor-sum decomposition $\mathcal{H} = \{1\} \oplus \mathcal{G}$; in the ANOVA decomposition of (1.1), $\eta_\emptyset \in \{1\}$ and $\sum_j \eta_j(x_j) + \sum_{j<k}\eta_{j,k}(x_j,x_k) + \cdots \in \mathcal{G}$. Writing $\eta = d + g$ with $g \in \mathcal{G}$ and $d$ a constant, (1.3) becomes

$$\frac{1}{n}\sum_{i=1}^n e^{-g(X_i)-d} + \int_{\mathcal{X}} \big\{g(x)+d\big\}\rho(x)dx + \frac{\lambda}{2}J(g). \tag{2.2}$$

Fixing $g(x)$, the $d$ that minimizes (2.2) is given by $e^d = n^{-1}\sum_{i=1}^n e^{-g(X_i)}$, noting that $\int_{\mathcal{X}}\rho(x)dx = 1$; in effect, (1.3) "normalizes" $\eta$ to satisfy $n^{-1}\sum_{i=1}^n e^{-\eta(X_i)} = 1$. Plugging this back into (2.2) and dropping terms not involving $g(x)$, one has a "profile" functional

$$\log\Big\{\frac{1}{n}\sum_{i=1}^n e^{-g(X_i)}\Big\} + \int_{\mathcal{X}} g(x)\rho(x)dx + \frac{\lambda}{2}J(g). \tag{2.3}$$

Without loss of inferential efficiency, one can minimize (1.3) in a space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \operatorname{span}\big\{R_J(Z_j,\cdot), j = 1,\ldots,q\big\}, \tag{2.4}$$

where $\{Z_j\}$ is a random subset of $\{X_i\}$; see Section 7.3. One has an expression

$$g(x) = \sum_\nu \breve{d}_\nu \phi_\nu(x) + \sum_j \breve{c}_j R_J(Z_j, x) = \breve{\boldsymbol{d}}^T \boldsymbol{\phi}(x) + \breve{\boldsymbol{c}}^T \breve{\boldsymbol{\xi}}(x) = \boldsymbol{c}^T \boldsymbol{\xi}(x), \qquad (2.5)$$

where $\{\phi_\nu\}$ is a basis of $\mathcal{N}_J \ominus \{1\}$ and $\breve{\xi}_j(x) = R_J(Z_j, x)$. Plugging (2.5) into (2.3), one has

$$\log\Big\{\frac{1}{n}\sum_{i=1}^n e^{-\boldsymbol{c}^T\boldsymbol{\xi}_i}\Big\} + \int_{\mathcal{X}} \boldsymbol{c}^T\boldsymbol{\xi}(x)\rho(x)dx + \frac{\lambda}{2}\boldsymbol{c}^T Q \boldsymbol{c}, \qquad (2.6)$$

where $\boldsymbol{\xi}_i = \boldsymbol{\xi}(X_i)$ and $Q = \mathrm{diag}(O, \breve{Q})$ for $\breve{Q}$ with $(j,k)$th entry $J(\breve{\xi}_j, \breve{\xi}_k) = R_J(Z_j, Z_k)$. Note that $\boldsymbol{b} = \int_{\mathcal{X}} \boldsymbol{\xi}(x)\rho(x)dx$ can be computed as sums of products of univariate integrals, for the $\phi_\nu$ and $R_J$ in the tensor-product cubic splines of Section 2.1 and for the choice of $\rho(x)$ as specified in Section 2.2; the involved univariate integrals and their products only need to be computed once, with only the sums to be performed for the updating of $\boldsymbol{b}$ when the $\theta$'s in (2.1) vary.

Fixing smoothing parameters, one can minimize (2.6) via the Newton iteration. Let $\tilde{\boldsymbol{c}}^T\boldsymbol{\xi}$ be the current iterate and define $\mu(f) = \sum_{i=1}^n u_i f(X_i)/\sum_{i=1}^n u_i$, where $u_i = e^{-\tilde{\boldsymbol{c}}^T\boldsymbol{\xi}_i}$. The Newton updating formula is seen to be $(V+\lambda Q)(\boldsymbol{c}-\tilde{\boldsymbol{c}}) = \boldsymbol{\mu} - \boldsymbol{b} - \lambda Q\tilde{\boldsymbol{c}}$, where $\boldsymbol{\mu} = \mu(\boldsymbol{\xi})$ and $V = \mu(\boldsymbol{\xi}\boldsymbol{\xi}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T$.

## 3. Smoothing Parameter Selection

With varying smoothing parameters that include the $\lambda$ in front of $J(\eta)$ and the $\theta$'s hidden in the reproducing kernel $R_J$ as seen in (2.1), the minimizer $\eta_\lambda$ of (1.3) provides a collection of estimates to choose from. The proper selection of smoothing parameters is crucial in practical estimation. For notational simplicity, $\lambda$ here represents all smoothing parameters, not just the $\lambda$ in front of $J(\eta)$.

In the examples of Jeon and Lin (2006), smoothing parameters were selected by a simple 5-fold cross-validation targeting the Kullback-Leibler loss. The normalizing constant $\int_{\mathcal{X}} e^{\eta_\lambda(x)}\rho(x)dx$ was needed, for which Monte Carlo integration was employed.

A certain delete-one cross-validation can be derived in the setting; it produces more consistent results at less computational cost than any $k$-fold cross-validation. The real challenge here is the normalizing constant, as the very motivation of (1.3) was to avoid integrals of form $\int_{\mathcal{X}} h(x)e^{\eta(x)}dx$; Monte-Carlo integration offers no relief as it is even less preferable to a regular quadrature for the purpose.

As an alternative to the Kullback-Leibler, consider a loss

$$\mathrm{LK}(\eta, \eta_\lambda) = \int_{\mathcal{X}} \big\{e^{(\eta-\eta_\lambda)(x)} - (\eta-\eta_\lambda)(x) - 1\big\}\rho(x)dx; \qquad (3.1)$$

note that $e^y - y - 1$ has a unique minimum at $y = 0$. Dropping terms not involving $\eta_\lambda$, one has

$$\mathrm{RLK}(\eta, \eta_\lambda) = \int_{\mathcal{X}} e^{-\eta_\lambda(x)} p(x)dx + \int_{\mathcal{X}} \eta_\lambda(x)\rho(x)dx, \tag{3.2}$$

where $p(x) = e^{\eta(x)}\rho(x)$; the first term may be estimated by a cross-validated sample mean, $n^{-1}\sum_{i=1}^n e^{-\eta_\lambda^{[i]}(X_i)}$, where $\eta_\lambda^{[i]}$ minimizes some delete-one version of (1.3).

Write $\eta = d + g = d + \boldsymbol{\xi}^T\boldsymbol{c}$ in (1.3) and denote its minimizer by $\eta_\lambda = \tilde\eta = \tilde d + \tilde g = \tilde d + \boldsymbol{\xi}^T\tilde{\boldsymbol{c}}$. Fixing $\tilde d$, consider the quadratic approximation of (1.3) at $\tilde\eta$ as a function of $\boldsymbol{c}$,

$$\frac{1}{n}\sum_{i=1}^n w_i\Big\{1 - \boldsymbol{\xi}_i^T(\boldsymbol{c} - \tilde{\boldsymbol{c}}) + \frac{1}{2}(\boldsymbol{c} - \tilde{\boldsymbol{c}})^T\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T(\boldsymbol{c} - \tilde{\boldsymbol{c}})\Big\} + \tilde d + \boldsymbol{b}^T\boldsymbol{c} + \frac{\lambda}{2}\boldsymbol{c}^T Q\boldsymbol{c}, \tag{3.3}$$

where $w_i = e^{-\tilde\eta(X_i)}$, $\boldsymbol{\xi}_i = \boldsymbol{\xi}(X_i)$, and $\boldsymbol{b} = \int_{\mathcal{X}}\boldsymbol{\xi}(x)\rho(x)dx$. The solution of (3.3) is of course $\tilde{\boldsymbol{c}}$, with $\tilde{\boldsymbol{c}} = A^{-1}\boldsymbol{d}$, where $A = n^{-1}\sum_{i=1}^n w_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T + \lambda Q$ and $\boldsymbol{d} = n^{-1}\sum_{i=1}^n w_i(1 + \tilde g_i)\boldsymbol{\xi}_i - \boldsymbol{b}$ for $\tilde g_i = \boldsymbol{\xi}_i^T\tilde{\boldsymbol{c}} = g_\lambda(X_i)$. Solving a delete-one version of (3.3),

$$\frac{1}{n}\sum_{j\neq i} w_j\Big\{1 - \boldsymbol{\xi}_j^T(\boldsymbol{c} - \tilde{\boldsymbol{c}}) + \frac{1}{2}(\boldsymbol{c} - \tilde{\boldsymbol{c}})^T\boldsymbol{\xi}_j\boldsymbol{\xi}_j^T(\boldsymbol{c} - \tilde{\boldsymbol{c}})\Big\} + \tilde d + \boldsymbol{b}^T\boldsymbol{c} + \frac{\lambda}{2}\boldsymbol{c}^T Q\boldsymbol{c},$$

one has $\tilde{\boldsymbol{c}}^{[i]} = \big(A - n^{-1}w_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\big)^{-1}\big(\boldsymbol{d} - n^{-1}w_i(1 + \tilde g_i)\boldsymbol{\xi}_i\big)$. We use $\tilde g_i^{[i]} = \boldsymbol{\xi}_i^T\tilde{\boldsymbol{c}}^{[i]}$ for $g_\lambda^{[i]}(X_i)$. Since

$$\big(A - n^{-1}w_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\big)^{-1} = A^{-1} + \frac{n^{-1}w_iA^{-1}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^TA^{-1}}{1 - n^{-1}w_i\boldsymbol{\xi}_i^TA^{-1}\boldsymbol{\xi}_i},$$

$\tilde\eta_i^{[i]} = \tilde\eta_i - a_i/(1-a_i)$ following straightforward algebra, where $a_i = n^{-1}w_i\boldsymbol{\xi}_i^TA^{-1}\boldsymbol{\xi}_i$. Substituting $n^{-1}\sum_i\tilde\eta_i^{[i]}$ for the first term in (3.2), one obtains a cross-validation score

$$V(\lambda) = \frac{1}{n}\sum_{i=1}^n e^{-\eta_\lambda(X_i)} + \int_{\mathcal{X}}\eta_\lambda(x)\rho(x)dx + \alpha\frac{1}{n}\sum_{i=1}^n e^{-\eta_\lambda(X_i)}\big(e^{a_i/(1-a_i)} - 1\big) \tag{3.4}$$

for $\alpha = 1$, which is the "pseudo likelihood" in (1.3) plus an extra term; in parallel settings such as (1.2), a fudge factor $\alpha \approx 1.4$ in front of the extra term in cross-validation scores proved to be effective in curbing severe undersmoothing with "bad" samples while suffering minimal performance degradation with "good" ones.

With a cross-validation score involving multiple smoothing parameters $(\lambda, \theta_\beta)$, one may use Algorithm 3.2 of Gu and Wahba (1991) to locate good initial values of $\theta_\beta$, which involves two passes of fixed-$\theta$ minimization, then perform quasi-Newton iteration to minimize the score. The initial value algorithm is highly effective, typically leaving only the "last 20% performance" for the quasi-Newton to pick up, thus it is often a good idea to skip the time-consuming quasi-Newton iteration. For the score $V(\lambda)$ of (3.4), however, one *must* skip the quasi-Newton step but for a different reason: *to salvage performance*; computational savings only count as afterthoughts in this situation. As a univariate function of $\lambda$, $V(\lambda)$ of (3.4) appears to be reasonably effective in tracking the loss, as is shown in the simulations of Section 5, but as a multivariate function of the $\theta$'s, it often loses track of the target it is designed to follow, delivering in the process poor performances or even outright disasters.

## 4. Squared Error Projection

To infer conditional independence structures characterized by selective elimination of interactions in log density $\eta$, one needs to assess the practical significance of ANOVA terms. The task resembles hypothesis testing, with $H_0 : \eta \in \mathcal{H}_0$ versus $H_a : \eta \in \mathcal{H}_0 \oplus \mathcal{H}_1$, say.

Lacking a sampling distribution in settings with infinite-dimensional nulls, the classical testing approach is of little help here. Instead, an approach based on the Kullback-Leibler geometry was developed in Gu (2004). For density estimation via (1.2), one calculates an estimate $\hat\eta \in \mathcal{H}_0 \oplus \mathcal{H}_1$, obtains its Kullback-Leibler projection $\tilde\eta \in \mathcal{H}_0$ by minimizing $\mathrm{KL}(\hat\eta, \eta)$ over $\eta \in \mathcal{H}_0$, then inspects the "entropy decomposition" $\mathrm{KL}(\hat\eta, \eta_u) = \mathrm{KL}(\hat\eta, \tilde\eta) + \mathrm{KL}(\tilde\eta, \eta_u)$, where $\eta_u = 0$ is the uniform distribution. When $\mathrm{KL}(\hat\eta, \tilde\eta)$ is only a small portion of $\mathrm{KL}(\hat\eta, \eta_u)$, say 2-3%, one loses little by cutting out $\mathcal{H}_1$.

The calculation of Kullback-Leibler projection also involves integrals of form $\int_{\mathcal{X}} h(x)e^{\eta(x)}dx$ as with (1.2), so is numerically impractical in the setting. As an alternative, one can consider $\tilde V(\hat\eta - \eta) = \int_{\mathcal{X}}(\hat\eta - \eta)^2(x)\rho(x)dx - \big\{\int_{\mathcal{X}}(\hat\eta - \eta)(x)\rho(x)dx\big\}^2$ for $\hat\eta \in \mathcal{H}_0 \oplus \mathcal{H}_1$, and calculate the squared error projection of $\hat\eta$ in $\mathcal{H}_0$ by minimizing $\tilde V(\hat\eta - \eta)$ over $\eta \in \mathcal{H}_0$; $\tilde V(\hat\eta - \eta)$ can be viewed as a proxy of the symmetrized Kullback-Leibler discrepancy, $\mathrm{KL}(\hat\eta, \eta) + \mathrm{KL}(\eta, \hat\eta)$, as discussed in Section 7.4. Note also that $\tilde V(\hat\eta - \eta)$ is invariant to the normalizing constant.

Let $\tilde\eta$ be the squared error projection of $\hat\eta$ in $\mathcal{H}_0$ and consider $A_{\tilde\eta,h}(\alpha) = \tilde V\big(\hat\eta - (\tilde\eta + \alpha h)\big)$ for $h \in \mathcal{H}_0$. Since $dA_{\tilde\eta,h}/d\alpha|_{\alpha=0} = 0$, $\tilde V(\hat\eta - \tilde\eta, h) = 0$, $\forall h \in \mathcal{H}_0$. The uniform distribution corresponds to $\eta_u = -\log\rho(x)$ and, when $\eta_u \in \mathcal{H}_0$, $\tilde V(\hat\eta - \tilde\eta, \tilde\eta - \eta_u) = 0$, so $\tilde V(\hat\eta - \eta_u) = \tilde V(\hat\eta - \tilde\eta) + \tilde V(\tilde\eta - \eta_u)$. When the ratio $\tilde V(\hat\eta - \tilde\eta)/\tilde V(\hat\eta - \eta_u)$ is small, one may safely cut out $\mathcal{H}_1$.

For $\rho(x)$ a product of marginal densities, as prescribed in Section 2.2, the calculations involved are sums of products of univariate integrals, and $\eta_u \in \mathcal{H}_0$ when $\mathcal{H}_0$ includes all the main effects.

## 5. Simulation Studies

The simulation studies presented here address various practical issues concerning the method being developed. Numerical experiments have been done on domains of dimensions three and five.

For the trivariate simulations on $[0,1]^3$, samples were taken from

$$f_3(x_1,x_2,x_3) \propto f_1(x_1-0.3x_3+0.1)f_1(x_2-0.2x_3+0.1)e^{-12.5(x_3-0.5)^2}I_{[0<x_1,x_2,x_3<1]}, \tag{5.1}$$

where $f_1(x) \propto e^{-50(x-0.3)^2} + 2e^{-50(x-0.7)^2}$ is the 1:2 mixture of $\mathcal{N}(0.3, 0.1^2)$ and $\mathcal{N}(0.7, 0.1^2)$. Note that $X_1 \perp X_2 | X_3$ here, so the correct model has log density of form $\eta = \eta_\emptyset + \eta_1 + \eta_2 + \eta_3 + \eta_{1,3} + \eta_{2,3}$.

For the five-dimensional simulations, we took $(X_2, X_3, X_4)^T \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} = (0.5)\mathbf{1}$ and $\Sigma^{-1} = \begin{pmatrix} 62 & -15 & 0 \\ -15 & 62 & -30 \\ 0 & -30 & 62 \end{pmatrix}$, $X_1 = Y_1 - 0.4X_2 - 0.1$, and $X_5 = Y_2 + 0.3X_4 - 0.1$, then truncated to $\mathcal{X} = [0,1]^5$, with $Y_1, Y_2 \sim f_1(y)$ the normal mixture given above, independent of $(X_2, X_3, X_4)^T$, and of each other. Note that $X_i \perp X_j |$(the rest) for $(i,j) = (1,3), (1,4), (1,5), (2,4), (2,5), (3,5)$, and the correct model has log density of form $\eta = \eta_\emptyset + \eta_1 + \eta_2 + \eta_3 + \eta_4 + \eta_5 + \eta_{1,2} + \eta_{2,3} + \eta_{3,4} + \eta_{4,5}$.

### 5.1. Empirical performance of cross-validation

To assess the practical performance of cross-validation, samples of size $n = 300$ were generated in the three-dimensional setting and tensor-product cubic splines were used to estimate the density under the correct model specification. A safe choice of $q = 100$ was used in (2.4); see Section 5.3 for simulations concerning the practical choice of $q$. For each of the one hundred replicates, three estimates were calculated via (1.3), two with the smoothing parameters $\lambda_v$ "minimizing" the cross-validation score (3.4) with $\alpha = 1, 1.4$, respectively, and the other with $\lambda_o$ minimizing the Kullback-Leibler loss

$$L(\lambda) = \mathrm{KL}(\eta, \eta_\lambda) = \int_{\mathcal{X}} (\eta - \eta_\lambda)(x) e^{\eta(x)} \rho(x) dx = \int_{\mathcal{X}} (\eta - \eta_\lambda)(x) p(x) dx;$$

only two passes of fixed-$\theta$ minimization were performed to locate $\lambda_v$ as noted in Section 3, but $\lambda_o$ did minimize $L(\lambda)$ as a multivariate function. Parallel results in the five-dimensional setting were also obtained for sample size $n = 600$ and $q = 100$. The standard $\mathrm{KL}(\eta, \eta_\lambda)$ loss was calculated for all estimates as the

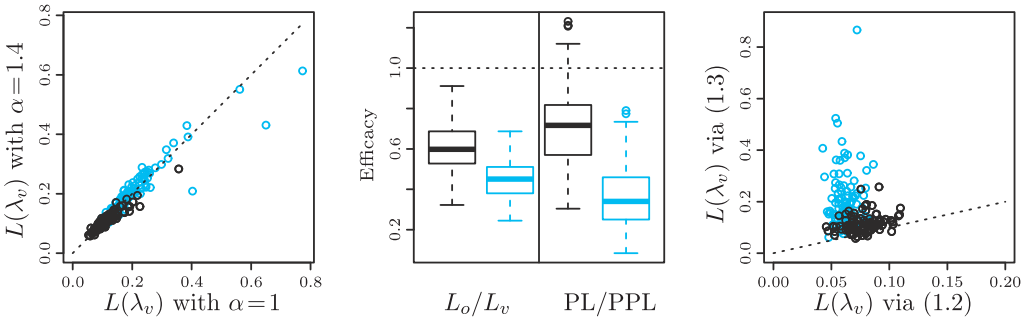


Figure 1. Performance of Cross-Validation. Left: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$. Right: $L(\lambda_v)$ via (1.2) versus $L(\lambda_v)$ via (1.3). Center: $L(\lambda_o)/L(\lambda_v)$ in the left half; $L(\lambda_v)$ via (1.2) over that via (1.3) in the right half. Results on $[0,1]^3$ with $n = 300$ are in solid and those on $[0,1]^5$ with $n = 600$ in faded.

performance measure, despite the use of the LK loss of (3.1) in the technical derivation of (3.4).

Plotted in the left frame of Figure 1 are $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$, suggesting a slight preference for $\alpha = 1.4$. The relative efficacy $L(\lambda_o)/L(\lambda_v)$ for $\alpha = 1.4$ is shown in the left half of the center frame in boxplots. The relative efficacy of $V(\lambda)$ in (3.4) is mediocre but acceptable; by using (1.3), we were ready to take a hit in performance in the first place, and the key here is to make things work reliably.

### 5.2. Comparisons against penalized likelihood

We now compare estimation via (1.3) with estimation via (1.2) in terms of performance and timing. For each of the samples, cross-validated estimates were calculated via (1.2) and (1.3), respectively, with the same $\{Z_j\}$ of size $q = 10n^{2/9}$ in (2.4) and the default $\alpha = 1.4$ in their respective cross-validation scores; the quasi-Newton step was skipped for the estimates via (1.2) to put things on an equal footing.

Shown in the right frame of Figure 1 are $L(\lambda_v)$ via (1.2) versus $L(\lambda_v)$ via (1.3), using one hundred replicates each with $n = 300$ on $[0,1]^3$, and with $n = 600$ on $[0,1]^5$; the ratios of $L(\lambda_v)$ via (1.2) over that via (1.3) are summarized in the right half of the center frame. As expected, (1.3) is no match to (1.2) in performance, but the degradation can still be measured.

To visually inspect the performances of the estimates, we pick the first replicate on $[0,1]^3$, which has $L(\lambda_v) = 0.1012$ via (1.3) and $L(\lambda_v) = 0.0867$ via (1.2). Shown in Figure 2 are the conditional densities $f(x_1, x_2|x_3 = 0.5)$ and $f(x_1, x_2|x_3 = 0.8)$; note that $f(x_1, x_2|x_3) = f(x_1|x_3)f(x_2|x_3)$ for the test density

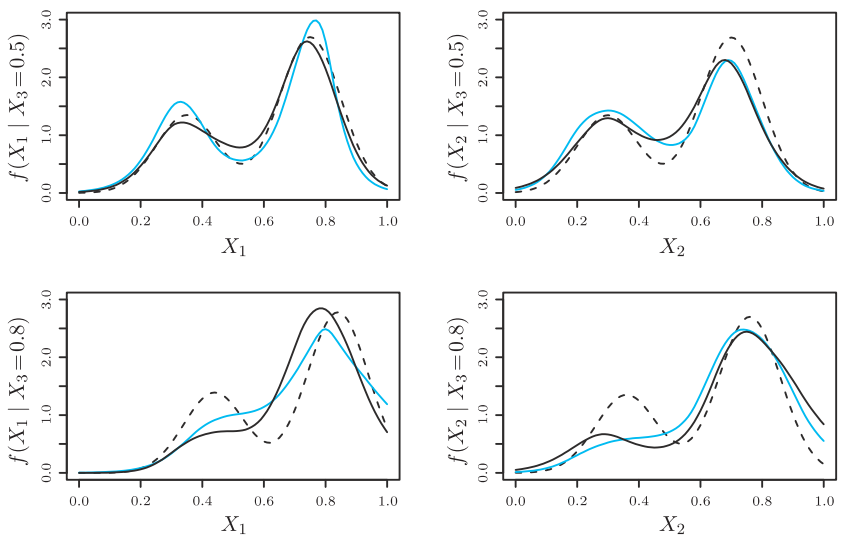


Figure 2. Estimation on $[0,1]^3$: A Fitted $f(x_1, x_2|x_3) = f(x_1|x_3)f(x_2|x_3)$. Estimates via (1.3) are in solid, those via (1.2) in faded, and the test density in dashed lines.

and the estimates. The mode of marginal density $f(x_3)$ is at $x_3 = 0.5$ and it is reassuring to see that $f(x_1, x_2|x_3 = 0.5)$ is estimated well, whereas less data are available near $x_3 = 0.8$ where the estimates are less accurate.

For the one hundred replicates on $[0,1]^3$ with $n = 300$ and $q = 36$, the estimates via (1.3) took 62.5 CPU seconds on a linux server, the estimates via (1.2) using a 2527-point quadrature took 296.4 CPU seconds, and (1.2) with a 3679-point quadrature took 405.9 CPU seconds. For the one hundred replicates on $[0,1]^5$ with $n = 600$ and $q = 42$, the estimates via (1.3) took 180.3 CPU seconds, the estimates via (1.2) using a 10063-point quadrature took 1839.7 CPU seconds, and (1.2) with a 17103-point quadrature took 3232.8 CPU seconds.

The computational savings here are moderate, probably not worth the performance degradation, but it is evident that the computation of (1.3) is roughly $O(nq^2)$ whereas that of (1.2) largely depends on the quadrature size. As the dimension goes up, adequate quadrature sizes quickly become astronomical, forcing (1.2) out of consideration. When (1.2) is numerically feasible, it is preferred to (1.3), though the latter remains a convenient exploratory tool, say for a quick check of the rough shape of $p(x_1|\text{the rest})$ or for an exploration of conditional independence structures.

### 5.3. Practical Choice of $q$

The computational cost increases with the dimension of $\mathcal{H}^*$ in (2.4) at a rate proportional to $q^2$, thus a small $q$ is preferable, while too small a $q$ could lead

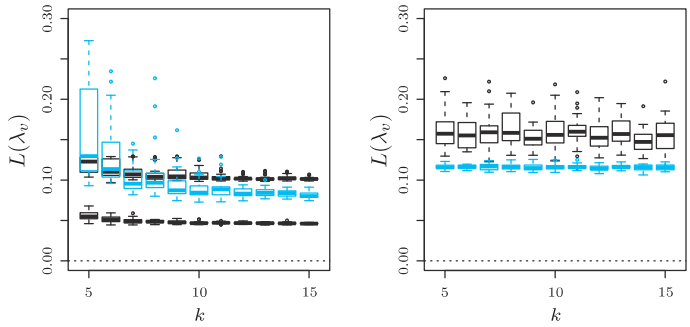


Figure 3. Effect of $q$ on Estimation Consistency. Boxplots of 30 $L(\lambda_v)$ for each of $q = kn^{2/9}$; from high to low, $n = 300$, 600, 1,200 on $[0,1]^3$ (Left) or $n = 600$, 1,200 on $[0,1]^5$ (Right).

to undesirable model bias. A proper balance is needed. As noted in Section 7.4, the optimal convergence rates $O_p\big(n^{-rp/(rp+1)}\big)$ are achieved as $qn^{-2/(rp+1)} \to \infty$ which, for tensor-product cubic splines (with $4 - \epsilon < r < 4$, $\forall \epsilon > 0$) and "supersmooth" $\eta_0$ (with $p = 2$), translates into $q \asymp n^{2/9+\epsilon}$, $\forall \epsilon > 0$. For practical estimation, one may use $q = kn^{2/9}$ with $k = 10$, as suggested by the following simulations.

In the three-dimensional setting, three samples were drawn from the test density, of sizes $n = 300$, 600, and $1,200$. For each sample and every $k$ on the grid 5(1)15, 30 different random subsets $\{Z_j\} \subset \{X_i\}$ of size $q = kn^{2/9}$ were generated to form 30 different $\mathcal{H}^*$, yielding 30 different cross-validated estimates under the correct model specification. The 30 $L(\lambda_v)$ for each $k$ are summarized in the boxplots in the left frame of Figure 3. For the $n = 300$ and $n = 1,200$ samples, things settled down around or before $k = 10$. The $n = 600$ sample appeared to be a "difficult" one, and things did not quite settle down until later. One factor at play here is that the random subsets $\{Z_j\}$ were also used for the estimation of the marginal densities that form the $\rho(x)$ function, thus (1.3) itself also varies with $\{Z_j\}$. We set $q = 10n^{2/9}$ as default in software implementation, which is appropriate for the "estimation" of $\rho(x)$; it can be easily overridden if a larger $q$ is desired.

Parallel results in the five-dimensional setting are shown in the right frame of Figure 3, for samples of sizes $n = 600$ and 1,200.

Note that the use of $q = 100 \approx 28(300)^{2/9} \approx 24(600)^{2/9}$ in the simulations of Section 5.1 kept to a minimum the effect of the choice of $\mathcal{H}^*$.

### 5.4. Squared error projection

For a quick check on the squared error projection of Section 4, one hundred replicates were drawn for each of $n = 300, 600$ in the three-dimensional setting.

Table 1. Quantiles of $\tilde{V}(\hat{\eta}-\tilde{\eta})/\tilde{V}(\hat{\eta}-\eta_u)$ for $n = 300, 600$ on $[0,1]^3$ and $n = 600$, 1,200 on $[0,1]^5$.

| | | 50% | 90% | 95% | 99% | 100% |
|---|---|---|---|---|---|---|
| $[0,1]^3$ | $n = 300$ | 0.0124 | 0.0342 | 0.0421 | 0.0644 | 0.1095 |
| | $n = 600$ | 0.0124 | 0.0332 | 0.0412 | 0.0657 | 0.0761 |
| $[0,1]^5$ | $n = 600$ | 0.0080 | 0.0187 | 0.0202 | 0.0306 | 0.0376 |
| | $n = 1200$ | 0.0045 | 0.0083 | 0.0107 | 0.0129 | 0.0130 |

Cross-validated estimates were calculated with the default $q = 10n^{2/9}$ and $\alpha = 1.4$ under a model specification involving all interactions,

$$\hat{\eta} = \eta_\emptyset + \eta_1 + \eta_2 + \eta_3 + \eta_{1,2} + \eta_{1,3} + \eta_{2,3} + \eta_{1,2,3},$$

and squared error projections were obtained in the correct model space, of the form

$$\tilde{\eta} = \eta_\emptyset + \eta_1 + \eta_2 + \eta_3 + \eta_{1,3} + \eta_{2,3}.$$

The ratios $\tilde{V}(\hat{\eta}-\tilde{\eta})/\tilde{V}(\hat{\eta}-\eta_u)$ are summarized in the first two rows of Table 1.

Parallel results in the five-dimensional setting were also obtained, for one hundred replicates each with $n = 600$ and 1,200, a model specification involving all two-way interactions,

$$\hat{\eta} = \eta_\emptyset+\eta_1+\eta_2+\eta_3+\eta_4+\eta_5+\eta_{1,2}+\eta_{1,3}+\eta_{1,4}+\eta_{1,5}+\eta_{2,3}+\eta_{2,4}+\eta_{2,5}+\eta_{3,4}+\eta_{3,5}+\eta_{4,5},$$

and projections of the form

$$\tilde{\eta} = \eta_\emptyset + \eta_1 + \eta_2 + \eta_3 + \eta_4 + \eta_5 + \eta_{1,2} + \eta_{2,3} + \eta_{3,4} + \eta_{4,5}.$$

The ratios $\tilde{V}(\hat{\eta}-\tilde{\eta})/\tilde{V}(\hat{\eta}-\eta_u)$ are summarized in the last two rows of Table 1.

The $[0,1]^3$, $n = 300$ line and the $[0,1]^5$, $n = 600$ line in Table 1 are based on the same replicates and the same $\{Z_j\}$ as the estimates shown in the right frame of Figure 1. The projection appears to be informative even if the estimation may be inaccurate.

## 6. Examples

The techniques being developed have been implemented in a suite of R functions included in the `gss` package. We now use the software tools to analyze a few data sets. Instead of the density estimates themselves which are less perceptible in high dimensions, we focus on the conditional independence structures, which can be depicted as undirected graphs.

### 6.1. Air pollution and road traffic

We first reanalyze an example in Jeon and Lin (2006). The data are found in the StatLib Datasets Archive at `http://lib.stat.cmu.edu/datasets/` under the heading `NO2`. It consists of a subset of 500 hourly observations collected at Alnabru in Oslo, Norway, between October 2001 and August 2003. The data are included in `gss` as a data frame `NO2` with elements `no2` ($NO_2$ concentration), `cars` (traffic volume), `temp` (temperature 2 meters above ground), `wind` (wind speed), `temp2` (temperature difference between 25 and 2 meters above ground), and `wind2` (wind direction).

To fit a joint density of the variables via (1.3), one can use:

```
library(gss); data(NO2); set.seed(5732)
fit <- ssden1(~(no2+cars+temp+wind+temp2+wind2)^2,data=NO2,
                                                  nbasis=100)
```

where all main effects and two-way interactions are included; `set.seed` ensures reproducible results and `nbasis` sets $q = 100$. To explore conditional independence structures, one can screen the significance of the interaction terms:

```
label <- fit$terms$labels[7:21]
ratio <- project(fit,include=label,drop1=TRUE)$ratio
```

where `label` lists the 15 interaction terms, and squared error projections are calculated with the terms removed one at a time. Normally, the argument `include` in the call to `project` specifies the terms included in $\mathcal{H}_0$ (all main effects are automatically included internally), but `drop1=TRUE` here asks for 15 projections each with 14 interaction terms in $\mathcal{H}_0$, with `ratio` containing the 15 "drop-one-term" $\tilde{V}(\hat{\eta}-\tilde{\eta})/\tilde{V}(\hat{\eta}-\eta_u)$ ratios labeled by the dropped interaction; these ratios indicate how irreplaceable the terms are in the fit, thus may be referred to as the "strengths" of the respective interaction terms. Putting things in a decreasing order:

```
rev(order(ratio))
```

one has $1, 8, 4, 3, 7, 2, 15, 6, 9, \ldots$. Projecting into spaces containing the first 8 and 9 terms, respectively, one has $\tilde{V}(\hat{\eta}-\tilde{\eta})/\tilde{V}(\hat{\eta}-\eta_u) = 3.3\%, 2.0\%$:

```
project(fit,include=label[c(1,8,4,3,7,2,15,6)])
project(fit,include=label[c(1,8,4,3,7,2,15,6,9)])
```

A graph depicting the 9 interactions in `label[c(1,8,4,3,7,2,15,6,9)]` is shown in Figure 4, where the labels on the edges mark the "strengths" of the links in the form of the "drop-one-term" $\tilde{V}(\hat{\eta}-\tilde{\eta})/\tilde{V}(\hat{\eta}-\eta_u)$ ratios. Given `(no2,cars)`, `(temp,wind)` and `(temp2,wind2)` are conditionally independent.

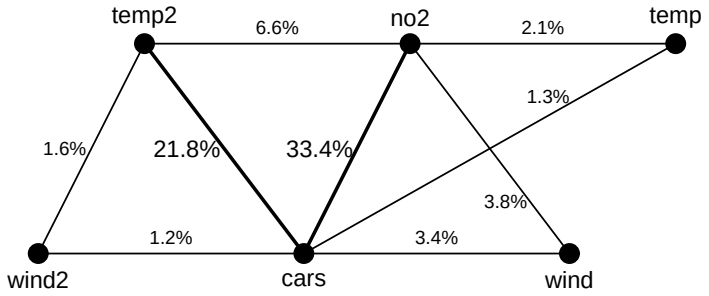


Figure 4. Graphical Model Fitted to `NO2` Data. The labels on the edges indicate how irreplaceable the interactions are in the fit.

### 6.2. Transcription factor association

Transcription factors play important roles in the study of gene expression. Some transcription factor association strength scores, normalized to be between 0 and 5.132242, were compiled by Ouyang, Zhou, and Wong (2009) for 12 transcription factors on 18936 genes, with the data available at

`http://www.pnas.org/content/suppl/2009/12/04/0904863106.DCSupplemental/SD2.txt`

One may read the data into R as a data frame:

```
SD2 <- read.table("SD2.txt",header=TRUE)
```

with elements `E2f1`, `Mycn`, `Zfx`, `Myc`, `Klf4`, `Tcfcp2l1`, `Esrrb`, `Nanog`, `Oct4`, `Sox2`, `Stat3`, and `Smad1`. A log density involving all main effects and two-way interactions was fitted to `SD2`:

```
set.seed(5732)
fit.sd2 <- ssden1(~(E2f1+Mycn+Zfx+Myc+Klf4+Tcfcp2l1+Esrrb
      +Nanog+Oct4+Sox2+Stat3+Smad1)^2,domain=domain,data=SD2)
```

where `domain` is a data frame with elements `E2f1=c(0,5.132242)`, `Mycn=c(0,5.132242)`, etc. specifying the domain $\mathcal{X} = [0, 5.132242]^{12}$ to be used in (1.3). Checking the "strengths" of the interaction terms and putting things in decreasing order:

```
lab.sd2 <- fit.sd2$terms$labels[-(1:12)]
r.sd2 <- project(fit.sd2,lab.sd2,TRUE)$ratio
rev(order(r.sd2))
```

one has $13, 3, 1, 2, 12, 46, 39, \ldots$. Projecting into spaces with the first 5 and 6 interactions, respectively, one has $\tilde{V}(\hat{\eta} - \tilde{\eta})/\tilde{V}(\hat{\eta} - \eta_u) = 3.3\%, 2.9\%$:

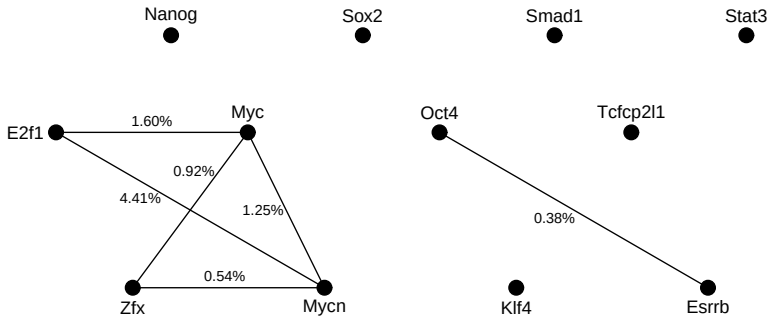


Figure 5. Graphical Model Fitted to SD2 Data. The labels on the edges indicate how irreplaceable the interactions are in the fit.

```
project(fit.sd2,lab.sd2[rev(order(r.sd2))[1:5]])
project(fit.sd2,lab.sd2[rev(order(r.sd2))[1:6]])
```

A graph illustrating the 6 terms in `lab.sd2[c(13,3,1,2,12,46)]` is shown in Figure 5. Apart from the first 6 terms, the rest of the terms all have "strengths" no better than 0.21%. The overall "weakness" of the interaction terms in this example suggest weak correlations among the transcription factors.

### 6.3. Protein signaling network

Measurements of 11 phosphorylated proteins/phospholipid components in 7466 human immune system cells were obtained by Sachs et al. (2005) and used to build maps of causal signaling pathways using Bayesian networks. We now fit a log density via (1.3) to the logorithms (base 10) of the measurements; the data are included in `gss` as a data frame `Sachs` with elements `praf`, `pmek`, `plcg`, `pip2`, `pip3`, `p44.42`, `pakts473`, `pka`, `pkc`, `p38`, and `pjnk`. A model involving all main effects and two-way interactions are fitted to the data, and the interactions are screened for their "strengths:"

```
data(Sachs); mn <- apply(Sachs[,-12],2,min);
mx <- apply(Sachs[,-12],2,max)
domain <- data.frame(rbind(mn,mx)); set.seed(5732)
fit.sachs <- ssden1(~(praf+pmek+plcg+pip2+pip3+p44.42+pakts473
                +pka+pkc+p38+pjnk)^2,domain=domain,data=Sachs)
lab.sachs <- fit.sachs$terms$labels[-(1:11)]
r.sachs <- project(fit.sachs,lab.sachs,TRUE)$ratio
```

Projecting into spaces with the first 14 and 22 terms, respectively:

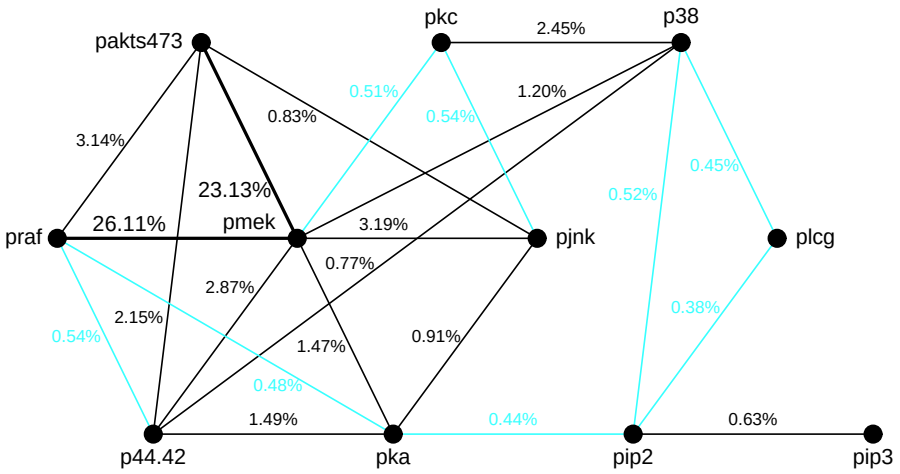


Figure 6. Graphical Model Fitted to `SACHS` Data. The labels on the edges indicate how irreplaceable the interactions are in the fit.

```
project(fit.sachs,lab.sachs[rev(order(r.sachs))[1:14]])
project(fit.sachs,lab.sachs[rev(order(r.sachs))[1:22]])
```

one has $\tilde{V}(\hat{\eta}-\tilde{\eta})/\tilde{V}(\hat{\eta}-\eta_u) = 6.8\%, 2.9\%$. The terms in `lab.sachs[rev(order(ratio))[1:22]]` are shown in the graph of Figure 6, where the 8 faded links have "strengths" no better than 0.54%. The structure appears to be much more complex than the one seen in Figure 5, with many more links irreplaceable.

Compared to the maps in Sachs et al. (2005), the undirected links here can only indicate associations, not causal pathways, though nodes "indirectly" associated do not have links between them. The results are obtained from a single fit following a top-down approach instead of model-averaging from hundreds of bottom-up estimates.

## 7. Asymptotic Convergence

We now analyze the asymptotic convergence properties of the minimizer $\hat{\eta}$ of (1.3). Denote by $e^{\eta_0(x)}\rho(x)$ the true density satisfying $\int e^{\eta_0(x)}\rho(x)dx = 1$, and take $V(f) = \int_{\mathcal{X}} f^2(x)\rho(x)dx$. Assuming $J(\eta_0) < \infty$, convergence rates will be established in terms of $V(\hat{\eta}-\eta_0)$. The analysis parallels that of (1.2) by Gu and Qiu (1993) using similar techniques, of which a slightly more polished presentation can be found in Gu (2002, Sec. 8.2).

The asymptotic analysis is based on an eigenvalue analysis of $V(f)$ with respect to $J(f)$, which is permitted by the following assumption.

**Assumption A.1** $V$ is completely continuous with respect to $V + J$.

Under A.1, there exist $\phi_\nu$ satisfying $V(\phi_\nu, \phi_\mu) = \delta_{\nu.\mu}$, $J(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu,\mu}$, and $0 \le \rho_\nu \uparrow \infty$, where $\delta_{\nu,\mu}$ is the Kronecker delta. The convergence rates are governed by the rate of growth of $\rho_\nu$.

**Assumption A.2** For some $r > 1$, $c > 0$, and $\nu$ sufficiently large, $\rho_\nu > c\nu^r$.

For the cubic spline of Section 2.1, $r = 4$. For tensor-product cubic splines, $4 - \epsilon < r < 4$, $\forall \epsilon > 0$.

### 7.1. Linear approximation

Consider the minimization of the quadratic functional

$$-\frac{1}{n}\sum_{i=1}^{n} e^{-\eta_0(X_i)}\eta(X_i) + \int_{\mathcal{X}} \eta(x)\rho(x)dx + \frac{1}{2}V(\eta - \eta_0) + \frac{\lambda}{2}J(\eta); \tag{7.1}$$

the minimizer $\tilde{\eta}$ of (7.1) is linear in data. Plugging $\eta(x) = \sum_\nu \eta_\nu \phi_\nu(x)$, $\eta_0(x) = \sum_\nu \eta_{\nu,0}\phi_\nu(x)$ into (7.1), where $\eta_\nu = V(\eta, \phi_\nu)$ and $\eta_{\nu,0} = V(\eta_0, \phi_\nu)$, one solves $\tilde{\eta}_\nu = (\beta_\nu + \eta_{\nu,0})/(1 + \lambda\rho_\nu)$, where $\beta_\nu = n^{-1}\sum_{i=1}^{n} e^{-\eta_0(X_i)}\phi_\nu(X_i) - \int_{\mathcal{X}} \phi_\nu(x)\rho(x)dx$.

Note that $E[\beta_\nu] = 0$ and $E[\beta_\nu^2] \le n^{-1}\int_{\mathcal{X}} \phi_\nu^2(x)e^{-\eta_0(x)}\rho(x)dx$. Following the lines of Gu (2002, Sec. 8.2.1), one has, as $n \to \infty$ and $\lambda \to 0$,

$$(V + \lambda J)(\tilde{\eta} - \eta_0) = O_p(\lambda + n^{-1}\lambda^{-1/r}); \tag{7.2}$$

a further assumption is needed, which holds when $\eta_0(x)$ is bounded from below as $V(\phi_\nu) = 1$.

**Assumption A.3** For some $c_3 < \infty$, $\int \phi_\nu^2(x)e^{-\eta_0(x)}\rho(x)dx < c_3$, $\forall \nu$.

### 7.2. Approximation error

Denote the minimizer of (1.3) by $\hat{\eta}$. Setting $\eta = \tilde{\eta} + \alpha g$ in (7.1) and $\eta = \hat{\eta} + \alpha g$ in (1.3), differentiating with respect to $\alpha$, and evaluating the derivatives at $\alpha = 0$, one has, $\forall g \in \mathcal{H}$,

$$-\frac{1}{n}\sum_{i=1}^{n} e^{-\eta_0(X_i)}g(X_i) + \int_{\mathcal{X}} g(x)\rho(x)dx + V(g, \tilde{\eta} - \eta_0) + \lambda J(g, \tilde{\eta}) = 0, \tag{7.3}$$

$$-\frac{1}{n}\sum_{i=1}^{n} e^{-\hat{\eta}(X_i)}g(X_i) + \int_{\mathcal{X}} g(x)\rho(x)dx + \lambda J(g, \hat{\eta}) = 0. \tag{7.4}$$

Subtracting (7.3) from (7.4) and setting $g = \hat{\eta} - \tilde{\eta}$, one has

$$\lambda J(\hat{\eta} - \tilde{\eta}) - \frac{1}{n}\sum_{i=1}^{n}\{e^{-\hat{\eta}(X_i)} - e^{-\tilde{\eta}(X_i)}\}(\hat{\eta} - \tilde{\eta})(X_i)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\{e^{-\tilde{\eta}(X_i)} - e^{-\eta_0(X_i)}\}(\hat{\eta} - \tilde{\eta})(X_i) + V(\hat{\eta} - \tilde{\eta}, \tilde{\eta} - \eta_0). \quad (7.5)$$

Further assumptions are needed to proceed; see Section 7.4 below for remarks.

**Assumption A.4** For some $0 < c_1 < c_2 < \infty$, and $\eta$ in a convex neighborhood of $\eta_0$ containing $\tilde{\eta}$, $\hat{\eta}$, and $\eta^*$ and $\hat{\eta}^*$ to be introduced in Section 7.3, $c_1 < e^{-\eta_0(x)+\eta(x)} < c_2$, $\forall x \in \mathcal{X}$.

**Assumption A.5** For some $c_4 < \infty$, $\int_{\mathcal{X}} \phi_\nu^2(x)\phi_\mu^2(x)e^{-\eta_0(x)}\rho(x)dx < c_4$, $\forall \nu, \mu$.

Under A.4, by the Mean Value Theorem, one has, for $c_1$ as in A.4 and some $c \in (c_1, c_2)$,

$$c_1\frac{1}{n}\sum_{i=1}^{n}e^{-\eta_0(X_i)}(\hat{\eta} - \tilde{\eta})^2(X_i)$$
$$\le -\frac{1}{n}\sum_{i=1}^{n}\{e^{-\hat{\eta}(X_i)} - e^{-\tilde{\eta}(X_i)}\}(\hat{\eta} - \tilde{\eta})(X_i), \quad (7.6)$$
$$-\frac{c}{n}\sum_{i=1}^{n}e^{-\eta_0(X_i)}(\hat{\eta} - \tilde{\eta})(X_i)(\tilde{\eta} - \eta_0)(X_i)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\{e^{-\tilde{\eta}(X_i)} - e^{-\eta_0(X_i)}\}(\hat{\eta} - \tilde{\eta})(X_i). \quad (7.7)$$

Under A.5, for $g(x) = \sum_\nu g_\nu\phi_\nu(x)$ and $h(x) = \sum_\nu h_\nu\phi_\nu(x)$, one has

$$\Big|\frac{1}{n}\sum_{i=1}^{n}e^{-\eta_0(X_i)}g(X_i)h(X_i) - V(g,h)\Big|$$
$$= \Big|\sum_\nu\sum_\mu g_\nu h_\mu\Big\{\frac{1}{n}\sum_i e^{-\eta_0(X_i)}\phi_\nu(X_i)\phi_\mu(X_i) - \int_{\mathcal{X}}\phi_\nu(x)\phi_\mu(x)\rho(x)dx\Big\}\Big|$$
$$\le \Big\{\sum_\nu\sum_\mu \frac{1}{1+\lambda\rho_\nu}\frac{1}{1+\lambda\rho_\mu}\Big\{\frac{1}{n}\sum_i e^{-\eta_0(X_i)}\phi_\nu(X_i)\phi_\mu(X_i)$$
$$- \int_{\mathcal{X}}\phi_\nu(x)\phi_\mu(x)\rho(x)dx\Big\}^2\Big\}^{1/2}\Big\{\sum_\nu\sum_\mu(1+\lambda\rho_\nu)(1+\lambda\rho_\mu)g_\nu^2h_\mu^2\Big\}^{1/2}$$
$$= (V+\lambda J)^{1/2}(g)(V+\lambda J)^{1/2}(h)O_p(n^{-1/2}\lambda^{-1/r}). \quad (7.8)$$

Substituting (7.6), (7.7), and (7.8) into (7.5), some manipulations yield, as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$,

$$(c_1 V + \lambda J)(\hat{\eta} - \tilde{\eta}) \le \big(|1-c| + o_p(1)\big)(V + \lambda J)^{1/2}(\hat{\eta} - \tilde{\eta})(V + \lambda J)^{1/2}(\tilde{\eta} - \eta_0)$$

which, in combination with (7.2), establishes the following.

**Theorem 2.** *Under Assumptions* A.1–A.5, *as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$,

$$(V + \lambda J)(\hat{\eta} - \eta_0) = O_p(\lambda + n^{-1}\lambda^{-1/r}).$$

### 7.3. Semiparametric approximation

The minimizer $\hat{\eta}$ of (1.3) in $\mathcal{H}$ is in general not computable. For practical applications, one needs computable approximations that do not sacrifice performance. We now consider the minimization of (1.3) in $\mathcal{H}^* = \mathcal{N}_J \oplus \text{span}\big\{R_J(Z_j, \cdot), j = 1, \ldots, q\big\}$, where $\{Z_j\}$ is a random subset of $\{X_i\}$.

Let $\eta^*$ be the projection of $\hat{\eta}$ in $\mathcal{H}^*$; $J(\eta^*, \hat{\eta} - \eta^*) = 0$. Setting $g = \hat{\eta} - \eta^*$ in (7.4), one has

$$-\frac{1}{n}\sum_{i=1}^{n} e^{-\hat{\eta}(X_i)}(\hat{\eta} - \eta^*)(X_i) + \int_{\mathcal{X}} (\hat{\eta} - \eta^*)(x)\rho(x)dx + \lambda J(\hat{\eta} - \eta^*, \hat{\eta}) = 0. \quad (7.9)$$

This can be rearranged as

$$\begin{aligned}\lambda J(\hat{\eta} - \eta^*) =& \frac{1}{n}\sum_i \big\{e^{-\hat{\eta}(X_i)} - e^{-\eta_0(X_i)}\big\}(\hat{\eta} - \eta^*)(X_i) \\ &+ \frac{1}{n}\sum_i e^{-\eta_0(X_i)}(\hat{\eta} - \eta^*)(X_i) - \int_{\mathcal{X}} (\hat{\eta} - \eta^*)(x)\rho(x)dx. \quad (7.10)\end{aligned}$$

The first term on the right-hand side of (7.10) is $\big(c + o_p(1)\big)V(\hat{\eta} - \eta_0, \hat{\eta} - \eta^*)$ for some $c$ by (7.8); the second term can be shown to be of the order $(V + \lambda J)^{1/2}(\hat{\eta} - \eta^*)O_p(n^{-1/2}\lambda^{-1/2r})$ using the same technique as in (7.8). For $h \in \mathcal{H} \ominus \mathcal{H}^*$, one has $V(h) = O_p(q^{-1/2}\lambda^{-1/r})(V + \lambda J)(h)$ (cf., Gu (2002, Lemma 8.5)), thus for $q^{-1/2}\lambda^{-1/r} \to 0$, $V(\hat{\eta} - \eta^*) = o_p\big(\lambda J(\hat{\eta} - \eta^*)\big)$. Substituting these into (7.10) one has, as $\lambda \to 0$ and $q^{1/2}\lambda^{1/r} \to \infty$,

$$(V + \lambda J)(\hat{\eta} - \eta^*) = O_p(\lambda + n^{-1}\lambda^{-1/r}). \quad (7.11)$$

Let $\hat{\eta}^*$ be the minimizer of (1.3) in $\mathcal{H}^*$. Setting $g = \hat{\eta} - \hat{\eta}^*$ in (7.4), one has

$$-\frac{1}{n}\sum_{i=1}^{n} e^{-\hat{\eta}(X_i)}(\hat{\eta} - \hat{\eta}^*)(X_i) + \int (\hat{\eta} - \hat{\eta}^*)(x)\rho(x) + \lambda J(\hat{\eta} - \hat{\eta}^*, \hat{\eta}) = 0. \quad (7.12)$$

Replacing $\hat{\eta}$ in (7.4) by $\hat{\eta}^*$ and setting $g = \hat{\eta}^* - \eta^*$, one has

$$-\frac{1}{n}\sum_{i=1}^{n} e^{-\hat{\eta}^*(X_i)}(\hat{\eta}^* - \eta^*)(X_i) + \int (\hat{\eta}^* - \eta^*)(x)\rho(x) + \lambda J(\hat{\eta}^* - \eta^*, \hat{\eta}^*) = 0. \quad (7.13)$$

Adding (7.12), (7.13) and subtracting (7.9), noting that $J(\hat{\eta}^* - \eta^*, \hat{\eta} - \eta^*) = 0$, some algebra yields

$$\begin{aligned}\lambda J(\hat{\eta}^* - \eta^*) &- \frac{1}{n}\sum_{i=1}^{n}\big\{e^{-\hat{\eta}^*(X_i)} - e^{-\eta^*(X_i)}\big\}(\hat{\eta}^* - \eta^*)(X_i) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\big\{e^{-\hat{\eta}(X_i)} - e^{-\eta^*(X_i)}\big\}(\hat{\eta}^* - \eta^*)(X_i). \quad (7.14)\end{aligned}$$

The left-hand side of (7.14) is no less than $\big(c_1 + o_p(1)\big)V(\hat{\eta}^* - \eta^*) + \lambda J(\hat{\eta}^* - \eta^*)$; the right-hand side is $\big(c + o_p(1)\big)V(\hat{\eta} - \eta^*, \hat{\eta}^* - \eta^*)$. These, in combination with (7.11) and Theorem 2, lead to the following.

**Theorem 3.** *Under Assumptions* A.1–A.5, *as* $\lambda \to 0$ *and* $q^{1/2}\lambda^{1/r} \to \infty$,

$$(V + \lambda J)(\hat{\eta}^* - \eta_0) = O_p(\lambda + n^{-1}\lambda^{-1/r}).$$

### 7.4. Remarks

The rate $O_p(\lambda + n^{-1}\lambda^{1/r})$ was established in (7.2) under the assumption that $J(\eta_0) = \sum_\nu \rho_\nu \eta_{\nu,0}^2 < \infty$, then propagated through subsequent analysis. When $\eta_0$ is "supersmooth" in the sense that $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ for some $p \in (1, 2]$, the rate can be improved to $O_p(\lambda^p + n^{-1}\lambda^{-1/r})$.

The optimal rate $O_p\big(n^{-rp/(rp+1)}\big)$ is achieved at $\lambda \asymp n^{-r/(rp+1)}$, which does satisfy $n\lambda^{2/r} \to \infty$ for $r > 1$ and $p \in [1, 2]$. For $q^{1/2}\lambda^{1/r} \to \infty$ with the optimal $\lambda$, one needs $qn^{-2/(rp+1)} \to \infty$.

Assumptions A.3 and A.4 hold automatically when $\eta_0(x)$ is bounded on $\mathcal{X}$ from below and above. Assumption A.5 is plausible in light of A.3, since $\phi_\mu(x)$ generally increases in "frequency" but not in magnitude as $\mu \to \infty$.

Define $\tilde{V}(f) = \int_{\mathcal{X}} \big\{f(x) - \int_{\mathcal{X}} f(x)\rho(x)dx\big\}^2 \rho(x)dx < V(f)$. $\tilde{V}(\eta - \eta_0)$ takes care of the normalizing constant, and rates in $V(\eta - \eta_0)$ imply rates in $\tilde{V}(\eta - \eta_0)$. For densities $p(x) \propto e^{\eta(x)}\rho(x)$ and $p'(x) \propto e^{\eta'(x)}\rho(x)$, the symmetrized Kullback-Leibler discrepancy between them is seen to be equal to $\int_{\mathcal{X}} \big\{(\eta - \eta')(x) - \int_{\mathcal{X}}(\eta - \eta')(x)\tilde{p}(x)dx\big\}^2 \tilde{p}(x)dx$, with $\tilde{p}(x) \propto e^{\tilde{\eta}(x)}\rho(x)$ for $\tilde{\eta}(x)$ a convex combination of $\eta(x)$ and $\eta'(x)$.

In the analysis of (1.2) by Gu and Qiu (1993), convergence rates were calculated in terms of $V^*(\eta - \eta_0) = \int_{\mathcal{X}}(\eta - \eta_0)^2(x)p(x)dx - \big\{\int_{\mathcal{X}}(\eta - \eta_0)(x)p(x)dx\big\}^2$,

where $p(x) = e^{\eta_0(x)}/\int_{\mathcal{X}} e^{\eta_0(x)}dx$ is the true density and $e^{\eta(x)}/\int_{\mathcal{X}} e^{\eta(x)}dx$ is the estimate; this agrees with $\tilde{V}(\eta - \eta_0)$ above for $\rho(x) = p(x)$.

## 8. Summary and Discussion

We have studied various aspects of an approach initiated by Jeon and Lin (2006) to nonparametric density estimation in high dimensions. Of primary concern is the effective smoothing parameter selection at a manageable computational cost, which makes or breaks the method. Also discussed are the "testing" of conditional independence structures and the asymptotic convergence of the estimates. Incorporating results from the theoretical, algorithmic, and simulation studies, a suite of R functions is made available for public use.

In our initial simulation studies, the performance of $V(\lambda)$ in (3.4) could be best described as hit-and-miss, highly unreliable even with the fudge factor. We were forced to explore numerous alternatives, but to no avail, as the only reliable alternative required the normalizing constant. The developments were set aside for a few years, until it was discovered, by sheer luck, that the score $V(\lambda)$, when "stopped" early, could be effective after all.

The formulation under study avoids the *numerical* burden of multi-dimensional integration in (1.2) at the cost of performance degradation, but it does not make the task of estimation any easier. Selective inclusion of ANOVA terms helps one to battle the curse of dimensionality, but large samples are needed for any estimation to be reliable in high dimensions. The squared error projection is similar to but different from a statistical test; see Gu (2004) for discussion concerning the similarities and philosophical differences.

In the simulation settings of Section 5, the savings in computational time may not be worth the performance degradation, and we do recommend the use of (1.2) when it is feasible. Adequate quadrature size grows quickly as the dimension goes up, however, and when the sample sizes are large enough to warrant density estimation in high dimensions, (1.2) quickly becomes infeasible numerically. Coupled with (3.4), (1.3) fills the void, and the squared error projection of Section 4 allows one to explore conditional independence structures even if the estimation may not be accurate.

## Acknowledgements

The authors thank Dr. Ping Ma for bringing to their attention the data set used in Section 6.2. Chong Gu's research was supported by the U.S. National Science Foundation under grant DMS-0705961. Yongho Jeon's research was supported by Basic Science Research Program of the National Research Foundation of Korea (2012-8-0666) funded by the Korean government.

## References

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68**, 337-404.

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58**, 255-277.

Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm. *J. Amer. Statist. Assoc.* **88**, 495-504.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer-Verlag.

Gu, C. (2004). Model diagnostics for smoothing spline ANOVA models. *Can. J. Statist.* **32**, 347-358.

Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: Theory. *Ann. Statist.* **21**, 217-234.

Gu, C. and Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.* **12**, 383-398.

Gu, C. and Wang, J. (2003). Penalized likelihood density estimation: Direct cross-validation and scalable approximation. *Statist. Sinica* **13**, 811-826.

Jeon, Y. and Lin, Y. (2006). An effective method for high dimensional log-density ANOVA estimation, with application to nonparametric graphical model building. *Statist. Sinica* **16**, 353-374.

Leonard, T. (1978). Density estimation, stochastic processes and prior information (with discussion). *J. Roy. Statist. Soc. Ser. B* **73**, 113-146.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Statist. Comput.* **9**, 363-379.

Ouyang, Z., Zhou Q. and Wong, W. H. (2009). ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **106**, 21521-21526. doi:10.1073/pnas.0904863106.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308 (5732)**, 523-529.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10**, 795-810.

Department of Statistics, Purdue University, West Lafayette, IN 47907, U.S.A.

E-mail: chong@stat.purdue.edu

Department of Applied Statistics, College of Business and Economics, Yonsei University, 50 Yonsei-Ro, Seodaemun-Gu, Seoul 120-749, Republic of Korea.

E-mail: yhjeon@yonsei.ac.kr

Verition Fund Management LLC, One American Lane, Greenwich, CT 06831, U.S.A.

E-mail: ylin22@gmail.com

(Received December 2011; accepted October 2012)