[1]

Sparse/Robust Estimation and Kalman Smoothing with Nonsmooth Log-Concave Densities: Modeling, Computation, and Theory

Aleksandr Y. Aravkin SARAVKIN@EOS.UBC.CA
Department of Earth and Ocean Sciences, University of British Columbia
Vancouver, BC, Canada

James V. Burke BURKE@MATH.WASHINGTON.EDU
Department of Mathematics, University of Washington
Seattle, WA, USA

Gianluigi Pillonetto GIAPI@DEI.UNIPD.IT
Department of Information Engineering, University of Padova
Padova, Italy

Editor:

Abstract

Piecewise linear quadratic (PLQ) penalties play a crucial role in many applications, including machine learning, robust statistical inference, sparsity promotion, and inverse problems such as Kalman smoothing. Well known examples of PLQ penalties include the ℓ_2, Huber, ℓ_1 and Vapnik losses. This paper builds on a dual representation for PLQ penalties known from convex analysis. We provide conditions that allow these losses to be interpreted as negative logs of true probability densities, and enable construction of non-smooth multivariate distributions with specified means and variances from simple scalar building blocks. The main contribution of this paper is a flexible statistical modelling framework for a variety of learning applications, where solutions to all models in this framework can be computed using interior point (IP) methods. IP methods solve nonsmooth optimization problems by working directly with smooth systems of equations characterizing the optimality of these problems. The efficiency of the IP approach depends on the structure of particular applications, and we consider the class of dynamic inverse problems using Kalman smoothing. This class comprises a wide variety of applications, where the aim is to reconstruct the state of a dynamical system with known process and measurement models starting from noisy output samples. In the classical case, Gaussian errors are assumed both in the process and measurement models for such problems. The extended framework allows arbitrary PLQ densities to be used, and the proposed IP approach solves the generalized Kalman smoothing problem while maintaining the linear complexity in the size of the time series, just as in the Gaussian case. This extends the computational efficiency of the Mayne-Fraser and Rauch-Tung-Striebel algorithms to a much broader nonsmooth setting, and includes many recently proposed robust and sparse smoothers as special cases.

Keywords: statistical modeling; nonsmooth optimization; robust inference; sparsity optimization; Kalman smoothing; interior point methods

[1] This work was in part financially supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (22R81254) and the Collaborative Research and Development Grant DNOISE II (375142-08). This research was carried out as part of the SINBAD II project with support from the following organizations: BG Group, BPG, BP, Chevron, Conoco Phillips, Petrobras, PGS, Total SA, and WesternGeco. The authors also wish to thank Bradley M. Bell for insightful discussion and helpful suggestions.

1. Introduction

Consider the classical problem of Bayesian parametric regression (MacKay, 1992; Roweis and Ghahramani, 1999) where the unknown $x \in \mathbb{R}^n$ is a random vector[1], with a prior distribution specified using a known invertible matrix $G \in \mathbb{R}^{n\times n}$ and known vector $\mu \in \mathbb{R}^n$ via

$$\mu = Gx + w \,, \tag{1.1}$$

where w is a zero mean vector with covariance Q. Let z denote a linear transformation of x contaminated with additive zero mean measurement noise v with covariance R,

$$z = Hx + v \,, \tag{1.2}$$

where $H \in \mathbb{R}^{\ell\times n}$ is a known matrix, while v and w are independent. It is well known that the (unconditional) minimum variance linear estimator of x, as a function of z, is the solution to the following optimization problem

$$\underset{x}{\operatorname{argmin}} \quad (z - Hx)^{\mathrm{T}} R^{-1} (z - Hx) + (\mu - Gx)^{\mathrm{T}} Q^{-1} (\mu - Gx) \,. \tag{1.3}$$

As we will show, (1.3) includes estimation problems arising in discrete-time dynamic linear systems which admit a state space representation (Anderson and Moore, 1979; Brockett, 1970). In this context, x is partitioned into N subvectors $\{x_k\}$, where each x_k represents the hidden system state at time instant k. For known data z, the classical Kalman smoother exploits the special structure of the matrices H, G, Q and R to compute the solution of (1.3) in $O(N)$ operations (Gelb, 1974). This procedure returns the minimum variance estimate of the state sequence $\{x_k\}$ when the additive noise in the system is assumed to be Gaussian.

In many circumstances, the estimator (1.3) performs poorly; put another way, quadratic penalization on model deviation is a bad model in many situations. For instance, it is not robust with respect to the presence of outliers in the data (Huber, 1981; Gao, 2008; Aravkin et al., 2011a; Farahmand et al., 2011) and may have difficulties in reconstructing fast system dynamics, e.g. jumps in the state values (Ohlsson et al., 2011). In addition, sparsity-promoting regularization is often used in order to extract a small subset from a large measurement or parameter vector which has greatest impact on the predictive capability of the estimate for future data. This sparsity principle permeates many well known techniques in machine learning and signal processing, including feature selection, selective shrinkage, and compressed sensing (Hastie and Tibshirani, 1990; Efron et al., 2004; Donoho, 2006). In these cases, (1.3) is often replaced by a more general model

$$\underset{x}{\operatorname{argmin}} \quad V\left(Hx - z; R\right) + W\left(Gx - \mu; Q\right) \tag{1.4}$$

where the loss V may be the ℓ_2-norm, the Huber penalty (Huber, 1981), Vapnik's ε-insensitive loss (used in support vector regression (Vapnik, 1998; Hastie et al., 2001)) or the hinge loss (leading to support vector classifiers (Evgeniou et al., 2000; Pontil and Verri, 1998; Schölkopf et al., 2000)). The regularizer W may be the ℓ_2-norm, the ℓ_1-norm (as in the LASSO (Tibshirani, 1996)), or a weighted combination of the two, yielding the elastic net procedure (Zou and Hastie, 2005).

[1] All vectors are column vectors, unless otherwise specified

Many learning algorithms using infinite-dimensional reproducing kernel Hilbert spaces as hypothesis spaces (Aronszajn, 1950; Saitoh, 1988; Cucker and Smale, 2001) boil down to solving finite-dimensional problems of the form (1.4) by virtue of the representer theorem (Wahba, 1998; Schölkopf et al., 2001).

These robust and sparse approaches can often be interpreted as placing non-Gaussian priors on w (or directly on x) and on the measurement noise v. The Bayesian interpretation of (1.4) has been extensively studied in the statistical and machine learning literature in recent years and probabilistic approaches used in the analysis of estimation and learning algorithms can be found e.g. in (Mackay, 1994; Tipping, 2001; Wipf et al., 2011). Non-Gaussian model errors and priors leading to a great variety of loss and penalty functions are also reviewed in (Palmer et al., 2006) using convex-type representations, and integral-type variational representations related to Gaussian scale mixtures.

In contrast to the above approaches, we consider a wide class of piecewise linear-quadratic (PLQ) functions and exploit their dual representation (Rockafellar and Wets, 1998). This class includes, among others, ℓ_2, ℓ_1, hinge loss, Huber and Vapnik losses, as we will show. We then establish conditions which allow these losses to be viewed as negative logs of true probability densities, ensuring that the vectors w and v come from true distributions. This in turn allows us to interpret the solution to the problem (1.4) as a MAP estimator when the loss functions V and W come from this subclass of PLQ penalties. This viewpoint allows statistical modelling using non-smooth penalties, and in particular we will show how multivariate densities with prescribed means and variances can be constructed using scalar PLQ penalties as building blocks.

In the second part of the paper, we derive the Karush-Kuhn-Tucker (KKT) system for problem (1.4), and introduce interior point (IP) methods, which directly attack the KKT system by working iteratively with smooth approximations. This allows a fundamentally smooth approach to many (non smooth) robust and sparse problems of interest to practitioners. Furthermore, we provide a theorem showing that IP methods solve (1.4) when V and W come from PLQ densities, subject to sufficient additional hypotheses, and describe implementation details for the entire class.

A concerted research effort has recently focused on the solution of regularized large scale inverse and learning problems, where computational costs and memory limitations are critical. This class of problems includes the popular kernel-based methods (Rasmussen and Williams, 2006; Schölkopf and Smola, 2001; Smola and Schölkopf, 2003), coordinate descent methods (Tseng and Yun, 2008; Lucidi et al., 2007; Dinuzzo, 2011) and decomposition techniques (Joachims, 1998; Lin, 2001; Lucidi et al., 2007), one of which is the widely used sequential minimal optimization algorithm for support vector machines (Platt, 1998). Other techniques are based on kernel approximations, e.g. using incomplete Cholesky factorization (Fine and Scheinberg, 2001), approximate eigen-decomposition (Zhang and Kwok, 2010) or truncated spectral representations (Pillonetto and Bell, 2007). Efficient interior point methods have been developed for ℓ_1-regularized problems (Kim et al., 2007), and for support vector machines (Ferris and Munson, 2003).

In contrast, general and efficient solvers for state space estimation problems of the form (1.4) are missing in the literature. The last part of this paper provides a contribution to fill this gap, specializing the general results to the dynamic case, and recovering the classical efficiency results of the least-squares formulation. In particular, we design new Kalman smoothers tailored for systems subject to noises coming from PLQ densities. Amazingly, it turns out that the IP method used in (Aravkin et al., 2011a) generalizes perfectly to the entire class of PLQ densities under a simple verifiable non-degeneracy condition. In practice, IP methods converge in a small number of iterations, and the effort per iteration depends on the structure of the underlying problem. We show that the IP iterations for all PLQ Kalman smoothing problems can be computed with a number of operations that scales linearly in N, as in the quadratic case. This theoretical foundation generalizes the results recently obtained in (Aravkin et al., 2011a,b; Farahmand et al., 2011; Ohlsson et al., 2011), framing them as particular cases of the general framework presented here.

The paper is organized as follows. In Section 2 we introduce the class of PLQ convex functions, and give sufficient conditions that allow us to interpret these functions as the negative logs of associated probability densities. In Section 3 we show how to construct multivariate densities with prescribed means and variances using scalar building blocks, with emphasis on densities corresponding to Huber and Vapnik penalties. In Section 4 we derive the KKT system for PLQ penalties from (Rockafellar and Wets, 1998), present a theorem that guarantees convergence of IP methods under appropriate hypotheses. In Section 5, we present the Kalman smoothing dynamic model, formulate Kalman smoothing with PLQ penaties, present the KKT system for the dynamic case, and show that IP iterations for PLQ smoothing preserve the classical computational effort results known for the Gaussian case. We present a numerical example in Section 6, and make some concluding remarks in Section 7. Section 8 serves as an appendix where supporting mathematical results and proofs are presented.

2. Piecewise Linear Quadratic Penalties and Densities

2.1 Preliminaries

We recall a few definitions from convex analysis, required to specify the domains of PLQ penalties. The reader is referred to (Rockafellar, 1970; Rockafellar and Wets, 1998) for more detailed reading.

- (Affine hull) Define the affine hull of any set $C \subset \mathbb{R}^n$, denoted by $\mathrm{aff}(C)$, as the smallest affine set that contains C.
- (Cone) For any set $C \subset \mathbb{R}^n$, denote by cone C the set $\{tr | r \in C, t \in \mathbb{R}_+\}$.
- (Domain) For $f(x) : \mathbb{R}^n \to \overline{\mathbb{R}} = \{\mathbb{R} \cup \infty\}$, $\mathrm{dom}(f) = \{x : f(x) < \infty\}$.
- (Polar Cone) For any cone $K \subset \mathbb{R}^m$, the polar of K is defined to be

$$K^\circ := \{r | \langle r, d\rangle \leq 0 \, \forall \, d \in K\}.$$

- (Horizon cone). Let $C \subset \mathbb{R}^n$ be a nonempty convex set. The horizon cone C^∞ is the convex cone of 'unbounded directions' for C, i.e. $d \in C^\infty$ if $C + d \subset C$.

2.2 PLQ densities

We now introduce the PLQ penalties and densities that are the focus of this paper. We begin with the dual representation of (Rockafellar and Wets, 1998), which is crucial to both establishing a statistical interpretation and to the development of a computational framework.

Definition 1 *(extended piecewise linear-quadratic penalties) (Rockafellar and Wets, 1998). Define* $\rho(U, M, b, B; \cdot) : \mathbb{R}^n \to \overline{\mathbb{R}}$ *as*

$$\rho(U, M, b, B; y) = \sup_{u \in U} \left\{ \langle u, b + By \rangle - \frac{1}{2} \langle u, Mu \rangle \right\} , \tag{2.1}$$

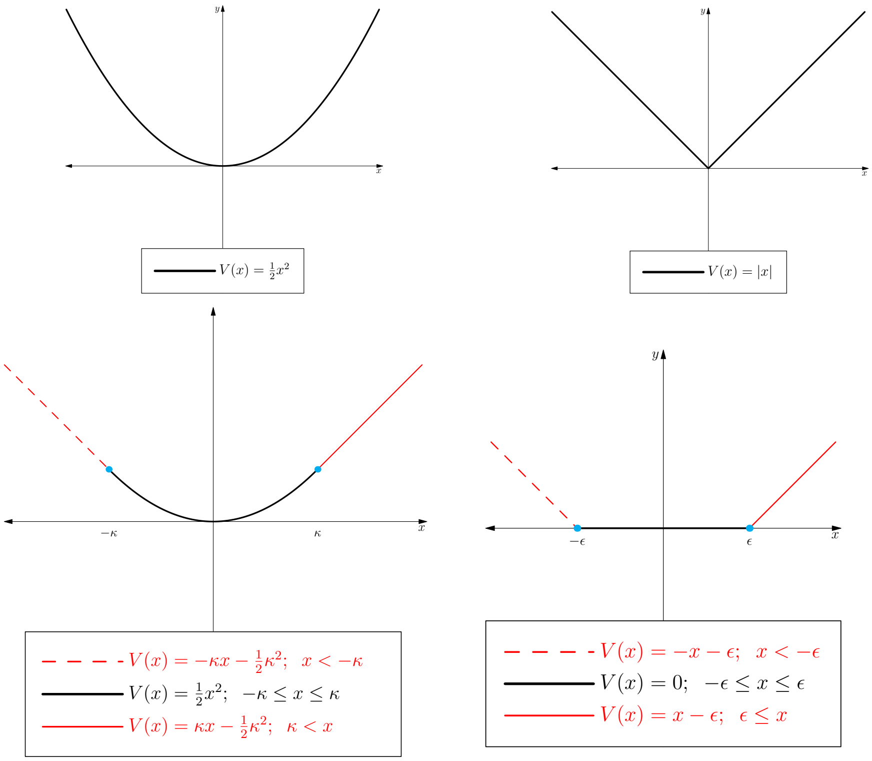

Figure 1: Scalar ℓ_2 (top left), ℓ_1 (top right), Huber (bottom left) and Vapnik (bottom right) Penalties

where $U \subset \mathbb{R}^m$ is a nonempty polyhedral set, $M \in \mathbb{R}^{m\times m}$ is a symmetric positive semidefinite matrix, and $b+By$ is an injective affine transformation with $B \in \mathbb{R}^{m\times n}$, so, in particular, $m \leq n$ and $\mathrm{null}(B) = \{0\}$.

Remark 2 *Taking $b = 0$ and $B = I$, we recover the basic piecewise linear-quadratic penalties characterized in (Rockafellar and Wets, 1998, Example 11.18). In particular,*

$$\mathrm{dom}[\rho(U,M,0,I;\cdot)] = [U^\infty \cap \mathrm{null}(M)]^\circ \ .$$

The following result characterizes the effective domain of ρ (see Appendix or (Aravkin, 2010) for proof).

Theorem 3 *Let ρ denote $\rho(U,M,B,b;y)$, and K denote $U^\infty \cap \mathrm{null}(M)$. Suppose $U \subset \mathbb{R}^m$ is a polyhedral set, $y \in \mathbb{R}^n$, $b \in K^\circ$, $M \in \mathbb{R}^{m\times m}$ is positive semidefinite, and $B \in \mathbb{R}^{m\times n}$ is injective. Then $(B^{\mathrm{T}}K)^\circ \subset \mathrm{dom}(\rho)$ and $[B^{\mathrm{T}}(K\cap -K)]^\perp = \mathrm{aff}[\mathrm{dom}(\rho)]$.*

Note that the functions ρ are still piecewise linear-quadratic. All of the examples previously mentioned can be represented in this way, as shown below.

Remark 4 (scalar examples) *ℓ_2, ℓ_1, elastic net, Huber, hinge, and Vapnik are all representable using the notation of Definition 1.*

1. *ℓ_2: Take $U = \mathbb{R}$, $M = 1$, $b = 0$, and $B = 1$. We obtain*

   $$\rho(y) = \sup_{u\in\mathbb{R}} \left\{uy - u^2/2\right\} \ .$$

   The function inside the sup *is maximized at $u = y$, hence $\rho(y) = \frac{1}{2}y^2$, see top left panel of Fig. 2.2.*

2. *ℓ_1: Take $U = [-1,1]$, $M = 0$, $b = 0$, and $B = 1$. We obtain*

   $$\rho(y) = \sup_{u\in[-1,1]} \{uy\} \ .$$

   The function inside the sup *is maximized by taking $u = \mathrm{sign}(y)$, hence $\rho(y) = |y|$, see top right panel of Fig. 2.2.*

3. *Elastic net, $\ell_2 + \lambda\ell_1$. This is a weighted sum of the previous two examples, and so must be in the class. Take*

   $$U = \mathbb{R}\times[-\lambda,\lambda],\ b = \begin{bmatrix}0\\0\end{bmatrix},\ M = \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix},\ B = \begin{bmatrix}1\\1\end{bmatrix} \ .$$

   This construction reveals the general calculus of PLQ addition.

4. *Huber: Take $U = [-\kappa,\kappa]$, $M = 1$, $b = 0$, and $B = 1$. We obtain $\rho(y) = \sup_{u\in U}\left\{uy - u^2/2\right\}$. We have the following cases:*

   (a) *If $y < -\kappa$, take $u = -\kappa$ to obtain $-\kappa y - \frac{1}{2}\kappa^2$.*

   (b) *If $-\kappa \le y \le \kappa$, take $u = y$ to obtain $\frac{1}{2}y^2$.*

   (c) *If $y > \kappa$, take $u = \kappa$ to obtain a contribution of $\kappa y - \frac{1}{2}\kappa^2$.*

   This is the Huber penalty, shown in the bottom left panel of Fig. 2.2.

5. *Hinge loss: Taking $B = 1$, $b = -\varepsilon$, $M = 0$ and $U = [0,1]$ we have*

   $$\rho(y) = \sup_{u\in U}\{(y-\varepsilon)u\} = (y-\varepsilon)_+.$$

   To verify this, just note that if $y < \varepsilon$, $u^* = 0$; otherwise $u^* = 1$.

6. *Vapnik loss is given by $(y-\varepsilon)_+ + (-y-\varepsilon)_+$. We immediately obtain its PLQ representation by taking*

   $$B = \begin{bmatrix}1\\-1\end{bmatrix},\ b = -\begin{bmatrix}\varepsilon\\\varepsilon\end{bmatrix},\ M = \begin{bmatrix}0 & 0\\0 & 0\end{bmatrix},\ U = [0,1]\times[0,1]$$

   to yield

   $$\rho(y) = \sup_{u\in U}\left\{\left\langle\begin{bmatrix}y-\varepsilon\\-y-\varepsilon\end{bmatrix}, u\right\rangle\right\} = (y-\varepsilon)_+ + (-y-\varepsilon)_+.$$

   The Vapnik penalty is shown in the bottom right panel of Fig. 2.2.

Note that the affine generalization (Definition 1) is already needed to express the Vapnik penalty and the elastic net, as both require summing together simpler PLQ penalties. Moreover, for both the elastic net and the Vapnik, the explicit representations are easily obtained from those of the summands. The constructions used are examples of a general pattern, as seen in the following remark.

Remark 5 *Let $\rho_1(y)$ and $\rho_2(y)$ be two PLQ penalties specified by U_i, M_i, b_i, B_i, for $i = 1,2$. Then the sum $\rho(y) = \rho_1(y) + \rho_2(y)$ is also a PLQ penalty, with*

$$U = U_1 \times U_2,\ M = \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix},\ b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},\ B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}\ .$$

In order to characterize PLQ penalties as negative logs of density functions, we need to ensure the integrability of said density functions. The function $\rho(y)$ is said to be *coercive* if $\lim_{\|y\|\to\infty} \rho(y) = \infty$, and coercivity turns out to be the key property to ensure integrability. The proof of this fact and the characterization of coercivity for PLQ penalties are the subject of the next two theorems (see Appendix for proofs).

Theorem 6 *(PLQ Integrability). Suppose $\rho(y)$ is coercive. Then the function* $\exp[-\rho(y)]$ *is integrable on* $\operatorname{aff}[\operatorname{dom}(\rho)]$ *with respect to the* $\dim(\operatorname{aff}[\operatorname{dom}(\rho)])$*-dimensional Lebesgue measure.*

Theorem 7 *A PLQ function ρ is coercive if and only if* $[B^{\mathrm{T}}\operatorname{cone}(U)]^\circ = \{0\}$.

Theorem 7 can be used to show the coercivity of familiar penalties.

Corollary 8 *The penalties ℓ_2, ℓ_1, elastic net, Vapnik, and Huber are all coercive.*

Proof We show all of these penalties satisfy the hypothesis of Theorem 7.

ℓ_2: $U = \mathbb{R}$ and $B = 1$, so $\left[B^{\mathrm{T}}\operatorname{cone}(U)\right]^\circ = \mathbb{R}^\circ = \{0\}$.

ℓ_1: $U = [-1,1]$, so $\operatorname{cone}(U) = \mathbb{R}$, and $B = 1$.

Elastic Net: In this case, $\operatorname{cone}(U) = \mathbb{R}^2$ and $B = \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$.

Huber: $U = [-\kappa, \kappa]$, so $\operatorname{cone}(U) = \mathbb{R}$, and $B = 1$.

Vapnik: $U = [0,1] \times [0,1]$, so $\operatorname{cone}(U) = \mathbb{R}^2_+$. $B = \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$, so $B^{\mathrm{T}}\operatorname{cone}(U) = \mathbb{R}$.

■

One can also show the coercivity of the above examples using their primal representations. However, our main objective is to pave the way for a modeling framework where multi-dimensional penalties can be constructed from simple building blocks and then solved by a uniform approach, using the dual representations alone.

We now define a family of distributions on $\mathbb{R}^n$ by interpreting piecewise linear quadratic functions ρ as negative logs of corresponding densities. Note that the support of the distributions is always contained in the affine set $\operatorname{aff}(\operatorname{dom}\rho)$, characterized in Th. 3.

Definition 9 *(Piecewise linear quadratic densities). Let $\rho(U,M,B,b;y)$ be any coercive extended piecewise linear quadratic function on $\mathbb{R}^n$. Define $\mathbf{p}(y)$ to be the following density on $\mathbb{R}^n$:*

$$\mathbf{p}(y) = \begin{cases} c^{-1}\exp\left[-\rho(y)\right] & y \in \operatorname{dom}\rho \\ 0 & \text{else}, \end{cases} \tag{2.2}$$

where

$$c = \left(\int_{y \in \operatorname{dom}\rho} \exp\left[-\rho(y)\right] dy\right),$$

and the integral is with respect to the $\dim(\operatorname{aff}[\operatorname{dom}(\rho)])$*-dimensional Lebesgue measure.*

PLQ densities are true densities on the affine hull of the domain of ρ. The proof of Theorem 6 can be easily adapted to show that they have moments of all orders.

3. Constructing PLQ densities

We make use of the following definitions. Given a sequence of column vectors $\{r_k\} = \{r_1,\ldots,r_N\}$ and matrices $\{\Sigma_k\} = \{\Sigma_1,\ldots,\Sigma_N\}$, we use the notation

$$\operatorname{vec}(\{r_k\}) = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix}, \ \operatorname{diag}(\{\Sigma_k\}) = \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \Sigma_N \end{bmatrix}.$$

In Definition 9, the PLQ densities are defined over $\mathbb{R}^n$. The moments of these densities depend in a nontrivial way on the choice of parameters b,B,U,M. In practice, we would like to be able to construct these densities to have prescribed means and variances. We now show how this can be done using scalar PLQ random variables as the building blocks. Suppose $y = \operatorname{vec}(\{y_k\})$ is a vector of independent (but not necessarily identical) PLQ random variables with mean 0 and variance 1. Denote by b_k,B_k,U_k,M_k the specification for the densities of y_k. To obtain the density of y, we need only take

$$\begin{aligned} U &= U_1 \times U_2 \times \cdots \times U_N \\ M &= \operatorname{diag}(\{M_k\}) \\ B &= \operatorname{diag}(\{B_k\}) \\ b &= \operatorname{vec}(\{b_k\}) . \end{aligned}$$

For example, the standard Gaussian distribution is specified by $U = \mathbb{R}^n$, $M = I$, $b = 0$, $B = I$, while the standard ℓ_1-Laplace (see (Aravkin et al., 2011a)) is specified by $U = [-1,1]^n$, $M = 0$, $b = 0$, $B = \sqrt{2}I$.

The random vector $\tilde{y} = Q^{1/2}(y+\mu)$ has mean μ and variance Q. If c is the normalizing constant for the density of y, then $c\det(Q)^{1/2}$ is the normalizing constant for the density of $\tilde{y}$.

Remark 10 *Note that only independence of the building blocks is required in the above result. This allows the flexibility to impose different PLQ densities on different errors in the model. Such flexibility may be useful for example when combining measurement data from different instruments, where some instruments may occasionally give bad data (with outliers), whereas others have nearly Gaussian error.*

We now show how to construct scalar building blocks with mean 0 and variance 1, i.e. how to compute the key normalizing constants for any PLQ penalty. To this aim, suppose $\rho(y)$ is a scalar PLQ penalty that is symmetric about 0. We would like to construct a density $\mathbf{p}(y) = \exp\left[-\rho(c_2 y)\right]/c_1$ to be a true density with unit variance, that is,

$$\frac{1}{c_1}\int \exp\left[-\rho(c_2 y)\right] dy = 1 \quad \text{and} \quad \frac{1}{c_1}\int y^2 \exp\left[-\rho(c_2 y)\right] dy = 1, \tag{3.1}$$

where the integrals are over $\mathbb{R}$. Using u-substitution, these equations become

$$c_1 c_2 = \int \exp\left[-\rho(y)\right] dy \quad \text{and} \quad c_1 c_2^3 = \int y^2 \exp\left[-\rho(y)\right] dy.$$

Solving this system yields

$$\begin{aligned} c_2 &= \sqrt{\int y^2 \exp\left[-\rho(y)\right] dy \Big/ \int \exp\left[-\rho(y)\right] dy} \\ c_1 &= \frac{1}{c_2}\int \exp\left[-\rho(y)\right] dy\,. \end{aligned}$$

These expressions can be used to obtain the normalizing constants for any particular ρ using simple integrals.

3.1 Huber Density

The scalar density corresponding to the Huber penalty is constructed as follows. Set

$$\mathbf{p_H}(y) = \frac{1}{c_1}\exp[-\rho_H(c_2 y)]\,, \tag{3.2}$$

where c_1 and c_2 are chosen as in (3.1). Specifically, we compute

$$\begin{aligned} \int \exp\left[-\rho_H(y)\right] dy &= 2\exp\left[-\kappa^2/2\right]\frac{1}{\kappa} + \sqrt{2\pi}[2\Phi(\kappa) - 1] \\ \int y^2 \exp\left[-\rho_H(y)\right] dy &= 4\exp\left[-K^2/2\right]\frac{1+\kappa^2}{\kappa^3} + \sqrt{2\pi}[2\Phi(\kappa) - 1]\,, \end{aligned}$$

where Φ is the standard normal cumulative density function. The constants c_1 and c_2 can now be readily computed.
To obtain the multivariate Huber density with variance Q and mean μ, let $U = [-\kappa, \kappa]^n$, $M = I$, $B = I$ any full rank matrix, and $b = 0$. This gives the desired density:

$$\mathbf{p_H}(y) = \frac{1}{c_1^n \det(Q^{1/2})}\exp\left[-\sup_{u\in U}\left\{\left\langle c_2 Q^{-1/2}(y-\mu), u\right\rangle - \frac{1}{2}u^{\mathrm{T}}u\right\}\right]. \tag{3.3}$$

3.2 Vapnik Density

The scalar density associated with the Vapnik penalty is constructed as follows. Set

$$\mathbf{p_V}(y) = \frac{1}{c_1}\exp\left[-\rho_V(c_2 y)\right]\,, \tag{3.4}$$

where the normalizing constants c_1 and c_2 can be obtained from

$$\int \exp\left[-\rho_V(y)\right] dy = 2(\varepsilon + 1)$$
$$\int y^2 \exp\left[-\rho_V(y)\right] dy = \frac{2}{3}\varepsilon^3 + 2(2 - 2\varepsilon + \varepsilon^2),$$

using the results in Section 3. Taking $U = [0,1]^{2n}$, the multivariate Vapnik distribution with mean μ and variance Q is

$$\mathbf{p_V}(y) = \frac{1}{c_1^n \det(Q^{1/2})} \exp\left[-\sup_{u\in U}\left\{\left\langle c_2 B Q^{-1/2}(y-\mu) - \varepsilon \mathbf{1}_{2n}, u\right\rangle\right\}\right], \tag{3.5}$$

where B is block diagonal with each block of the form $B = \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$, and $\mathbf{1}_{2n}$ is a column vector of 1's of length $2n$.

4. Optimization with PLQ penalties

We now return to the estimation problem (1.4) where the functions V and W are to be taken from the class of PLQ penalties. In the previous sections, PLQ penalties which are also negative log likelihoods of true densities were characterized using their dual representation. As we have seen, the scope of such densities is extremely broad, and, moreover, these densities can easily be constructed to possess specified moment properties. In this section, we expand on their utility by showing that the resulting estimation problems (1.4) can be solved with high accuracy using standard techniques from numerical optimization. We exploit the dual representation for the class of PLQ penalties (Rockafellar and Wets, 1998) to explicitly construct the Karush-Kuhn-Tucker (KKT) conditions for a wide variety of model problems of the form (1.4). Working with these systems opens the door to using a wide variety of numerical methods for convex quadratic programming to solve (1.4).

Let $\rho(U_v, M_v, b_v, B_v; y)$ and $\rho(U_w, M_w, b_w, B_w; y)$ be two PLQ penalties and define

$$V(v; R) := \rho(U_v, M_v, b_v, B_v; R^{-1/2} v) \tag{4.1}$$

and

$$W(w; Q) := \rho(U_w, M_w, b_w, B_w; Q^{-1/2} w). \tag{4.2}$$

Then (1.4) becomes

$$\min_{y\in\mathbb{R}^n} \rho(U, M, b, B; y), \tag{4.3}$$

where

$$U := U_v \times U_w,\ M := \begin{bmatrix} M_v & 0 \\ 0 & M_w \end{bmatrix},\ b := \begin{pmatrix} b_v - B_v R^{-1/2} z \\ b_w - B_w Q^{-1/2} \mu \end{pmatrix},$$

and

$$B := \begin{bmatrix} B_v R^{-1/2} H \\ B_w Q^{-1/2} G \end{bmatrix}.$$

Moreover, the hypotheses in (1.1), (1.2), (1.4), and (2.1) imply that the matrix B in (4.3) is injective. Indeed, $By = 0$ if and only if $B_w Q^{-1/2} G y = 0$, but, since G is nonsingular and B_w is injective, this implies that $y = 0$. That is, $\mathrm{nul}(B) = \{0\}$. Consequently, the objective in (4.3) takes the form of a

PLQ penalty function (2.1). In particular, if (4.1) and (4.2) arise from PLQ densities (definition 9), then the solution to problem (4.3) is the MAP estimator in the statistical model (1.1)-(1.2).

To simplify the notational burden, in the remainder of this section we work with (4.3) directly and assume that the defining objects in (4.3) have the dimensions specified in (2.1);

$$U \in \mathbb{R}^m,\ M \in \mathbb{R}^{m\times m},\ b \in \mathbb{R}^m,\ \text{and}\ B \in \mathbb{R}^{m\times n}. \tag{4.4}$$

The Lagrangian (Rockafellar and Wets, 1998)[Example 11.47] for problem (4.3) is given by

$$L(y,u) = b^{\mathrm{T}}u - \frac{1}{2}u^{\mathrm{T}}Mu + u^{\mathrm{T}}By\ .$$

By assumption U is polyhedral, and so can be specified to take the form

$$U = \{u : A^{\mathrm{T}}u \le a\}\ , \tag{4.5}$$

where $A \in \mathbb{R}^{m\times \ell}$. Using this reprsentation for U, the optimality conditions for (4.3) (Rockafellar, 1970; Rockafellar and Wets, 1998) are

$$\begin{aligned}
0 &= B^{\mathrm{T}}u \\
0 &= b + By - Mu - Aq \\
0 &= A^{\mathrm{T}}u + s - a \\
0 &= q_i s_i\ \ i = 1,\ldots,\ell\ ,\ q,s \ge 0\ ,
\end{aligned} \tag{4.6}$$

where the non-negative slack variable s is defined by the third equation in (4.6). The non-negativity of s implies that $u \in U$. The equations $0 = q_i s_i\ \ i = 1,\ldots,\ell$ in (4.6) are known as the complementarity conditions. By convexity, solving the problem (4.3) is equivalent to satisfying (4.6). There is a vast optimization literature on working directly with the KKT system. In particular, interior point (IP) methods (Kojima et al., 1991; Nemirovskii and Nesterov, 1994; Wright, 1997) can be employed. In the Kalman filtering/smoothing application, IP methods have been used to solve the KKT system (4.6) in a numerically stable and efficient manner, see e.g. (Aravkin et al., 2011b). Remarkably, the IP approach used in (Aravkin et al., 2011b) generalizes to the entire PLQ class. For Kalman filtering and smoothing, the computational efficiency is also preserved (see Section 5. Here, we show the general development for the entire PLQ class using standard techniques from the IP literature.

Let U, M, b, B, and A be as defined in (2.1) and (4.5), and let $\tau \in (0,+\infty]$. We define the τ *slice of the strict feasibility region for* (4.6) to be the set

$$\mathscr{F}_+(\tau) = \left\{ (s,q,u,y)\ \middle|\ \begin{array}{c} 0 < s,\ 0 < q,\ s^{\mathrm{T}}q \le \tau,\ \text{and} \\ (s,q,u,y)\ \text{satisfy the affine equations in (4.6)} \end{array} \right\}\ ,$$

and the *central path for* (4.6) to be the set

$$\mathscr{C} := \left\{ (s,q,u,y)\ \middle|\ \begin{array}{c} 0 < s,\ 0 < q,\ \gamma = q_i s_i\ \ i = 1,\ldots,\ell,\ \text{and} \\ (s,q,u,y)\ \text{satisfy the affine equations in (4.6)} \end{array} \right\}\ .$$

For simplicity, we define $\mathscr{F}_+ := \mathscr{F}_+(+\infty)$. The basic strategy of a primal-dual IP method is to follow the central path to a solution of (4.6) as $\gamma \downarrow 0$ by applying a predictor-corrector damped

Newton method to the function mapping $\mathbb{R}^\ell \times \mathbb{R}^\ell \times \mathbb{R}^m \times \mathbb{R}^n$ to itself given by

$$F_\gamma(s,q,u,y) = \begin{bmatrix} s + A^{\mathrm{T}}u - a \\ D(q)D(s)\mathbf{1} - \gamma\mathbf{1} \\ By - Mu - Aq + b \\ B^{\mathrm{T}}u \end{bmatrix} , \tag{4.7}$$

where $D(q)$ and $D(s)$ are diagonal matrices with vectors q, s on the diagonal.

Theorem 11 *Let U, M, b, B, and A be as defined in* (2.1) *and* (4.5). *Given $\tau > 0$, let $\mathscr{F}_+$, $\mathscr{F}_+(\tau)$, and $\mathscr{C}$ be as defined above. If*

$$\mathscr{F}_+ \neq \emptyset \quad \textit{and} \quad \mathrm{null}(M) \cap \mathrm{null}(A^{\mathrm{T}}) = \{0\}, \tag{4.8}$$

then the following statements hold.

(i) $F_\gamma^{(1)}(s,q,u,y)$ *is invertible for all* $(s,q,u,y) \in \mathscr{F}_+$.

(ii) Define $\widehat{\mathscr{F}_+} = \{(s,q) \mid \exists (u,y) \in \mathbb{R}^m \times \mathbb{R}^n \text{ s.t. } (s,q,u,y) \in \mathscr{F}_+\}$. *Then for each* $(s,q) \in \widehat{\mathscr{F}_+}$ *there exists a unique* $(u,y) \in \mathbb{R}^m \times \mathbb{R}^n$ *such that* $(s,q,u,y) \in \mathscr{F}_+$.

(iii) The set $\mathscr{F}_+(\tau)$ *is bounded for every* $\tau > 0$.

(iv) For every $g \in \mathbb{R}^\ell_{++}$, *there is a unique* $(s,q,u,y) \in \mathscr{F}_+$ *such that* $g = (s_1q_1, s_2q_2, \ldots, s_\ell q_\ell)^{\mathrm{T}}$.

(v) For every $\gamma > 0$, *there is a unique solution* $[s(\gamma), q(\gamma), u(\gamma), y(\gamma)]$ *to the equation* $F_\gamma(s,q,u,y) = 0$. *Moreover, these points form a differentiable trajectory in* $\mathbb{R}^\nu \times \mathbb{R}^\nu \times \mathbb{R}^m \times \mathbb{R}^n$. *In particular, we may write*

$$\mathscr{C} = \{[s(\gamma), q(\gamma), u(\gamma), y(\gamma)] \mid \gamma > 0\} .$$

(vi) The set of cluster points of the central path as $\gamma \downarrow 0$ *is non-empty, and every such cluster point is a solution to* (4.6).

Please see the Appendix for proof. Theorem 11 shows that if the conditions (4.8) hold, then IP techniques can be applied to solve the problem (4.3). In all of the applications we consider, the condition $\mathrm{null}(M) \cap \mathrm{null}(A^{\mathrm{T}}) = \{0\}$ is easily verified. For example, in the setting of (4.3) with

$$U_v = \{u \mid A_v u \leq a_v\} \quad \text{and} \quad U_w = \{u \mid A_w u \leq b_w\} \tag{4.9}$$

this condition reduces to

$$\mathrm{null}(M_v) \cap \mathrm{null}(A_v^{\mathrm{T}}) = \{0\} \quad \text{and} \quad \mathrm{null}(M_w) \cap \mathrm{null}(A_w^{\mathrm{T}}) = \{0\}. \tag{4.10}$$

Corollary 12 *The densities corresponding to* ℓ_1, ℓ_2, *Huber, and Vapnik penalties all satisfy hypothesis* (4.10).

Proof We verify that $\mathrm{null}(M) \cap \mathrm{null}(A^{\mathrm{T}}) = 0$ for each of the four penalties. In the ℓ_2 case, M has full rank. For the ℓ_1, Huber, and Vapnik penalties, the respective sets U are bounded, so $U^\infty = \{0\}$. ■

On the other hand, the condition $\mathscr{F}_+ \neq \emptyset$ is typically more difficult to verify. We show how this is done for two sample cases from class (1.4), where the non-emptiness of $\mathscr{F}_+$ is established by constructing an element of this set. Such constructed points are useful for initializing the interior point algorithm.

4.1 $\ell_1 - \ell_2$:

Suppose $V(v;R) = \left\|R^{-1/2}v\right\|_1$ and $W(w;Q) = \frac{1}{2}\left\|Q^{-1/2}w\right\|_2^2$. In this case

$$U_v = [-\mathbf{1}_m, \mathbf{1}_m],\ M_v = 0_{m\times m},\ b_v = 0_m,\ B_v = I_{m\times m},$$
$$U_w = \mathbb{R}^n,\ M_w = I_{n\times n},\ b_w = 0_n,\ B_w = I_{n\times n},$$

and $R \in \mathbb{R}^{m\times m}$ and $Q \in \mathbb{R}^{n\times n}$ are symmetric positive definite covariance matrices. Following the notation of (4.3) we have

$$U = [-\mathbf{1}, \mathbf{1}] \times \mathbb{R}^n,\ M = \begin{bmatrix} 0_{m\times m} & 0 \\ 0 & I_{n\times n} \end{bmatrix},\ b = \begin{pmatrix} -R^{-1/2}z \\ -Q^{-1/2}\mu \end{pmatrix},\ B = \begin{bmatrix} R^{-1/2}H \\ Q^{-1/2}G \end{bmatrix}.$$

The specification of U in (4.5) is given by

$$A^{\mathrm{T}} = \begin{bmatrix} I_{m\times m} & 0_{n\times n} \\ -I_{m\times m} & 0_{n\times n} \end{bmatrix} \text{ and } a = \begin{pmatrix} \mathbf{1} \\ -\mathbf{1} \end{pmatrix}.$$

Clearly, the condition $\mathrm{null}(M) \cap \mathrm{null}(A^{\mathrm{T}}) = \{0\}$ in (4.8) is satisfied. Hence, for Theorem 11 to apply, we need only check that $\mathscr{F}_+ \neq \emptyset$. This is easily established by noting that $(s,q,u,y) \in \mathscr{F}_+$, where

$$u = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ y = G^{-1}\mu,\ s = \begin{pmatrix} \mathbf{1} \\ \mathbf{1} \end{pmatrix},\ q = \begin{pmatrix} \mathbf{1} + [R^{-1/2}(Hy - z)]_+ \\ \mathbf{1} - [R^{-1/2}(Hy - z)]_- \end{pmatrix},$$

where, for $g \in \mathbb{R}^\ell$, g_+ is defined componentwise by $g_{+(i)} = \max\{g_i, 0\}$ and $g_{-(i)} = \min\{g_i, 0\}$.

4.2 Vapnik – Huber:

Suppose that $V(v;R)$ and $W(w;Q)$ are as in (4.1) and (4.2), respectively, with V a Vapnik penalty and W a Huber penalty:

$$U_v = [0, \mathbf{1}_m] \times [0, \mathbf{1}_m],\ M_v = 0_{2m\times 2m},\ b_v = -\begin{pmatrix} \varepsilon\mathbf{1}_m \\ \varepsilon\mathbf{1}_m \end{pmatrix},\ B_v = \begin{bmatrix} I_{m\times m} \\ -I_{m\times m} \end{bmatrix}$$
$$U_w = [-\kappa\mathbf{1}_n, \kappa\mathbf{1}_n],\ M_w = I_{n\times n},\ b_w = 0_n,\ B_w = I_{n\times n}\ ,$$

and $R \in \mathbb{R}^{m\times m}$ and $Q \in \mathbb{R}^{n\times n}$ are symmetric positive definite covariance matrices. Following the notation of (4.3) we have

$$U = ([0, \mathbf{1}_m] \times [0, \mathbf{1}_m]) \times [-\kappa\mathbf{1}_n, \kappa\mathbf{1}_n],\ M = \begin{bmatrix} 0_{2m\times 2m} & 0 \\ 0 & I_{n\times n} \end{bmatrix},$$
$$b = -\begin{pmatrix} \varepsilon\mathbf{1}_m + R^{-1/2}z \\ \varepsilon\mathbf{1}_m - R^{-1/2}z \\ Q^{-1/2}\mu \end{pmatrix},\ B = \begin{bmatrix} R^{-1/2}H \\ -R^{-1/2}H \\ Q^{-1/2}G \end{bmatrix}.$$

The specification of U in (4.5) is given by

$$A^{\mathrm{T}} = \begin{bmatrix} I_{m\times m} & 0 & 0 \\ -I_{m\times m} & 0 & 0 \\ 0 & I_{m\times m} & 0 \\ 0 & -I_{m\times m} & 0 \\ 0 & 0 & I_{n\times n} \\ 0 & 0 & -I_{n\times n} \end{bmatrix} \text{ and } a = \begin{pmatrix} \mathbf{1}_m \\ 0_m \\ \mathbf{1}_m \\ 0_m \\ \kappa\mathbf{1}_n \\ \kappa\mathbf{1}_n \end{pmatrix}.$$

Since $\mathrm{null}(A^{\mathrm{T}}) = \{0\}$, the condition $\mathrm{null}(M) \cap \mathrm{null}(A^{\mathrm{T}}) = \{0\}$ in (4.8) is satisfied. Hence, for Theorem 11 to apply, we need only check that $\mathscr{F}_+ \neq \emptyset$. We establish this by constructing an element (s,q,u,y) of $\mathscr{F}_+$. For this, let

$$u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}, \; s = \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \end{pmatrix}, \; q = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \\ q_6 \end{pmatrix},$$

and set

$$y = 0_n,\; u_1 = u_2 = \frac{1}{2}\mathbf{1}_\ell,\; u_3 = 0_n,\; s_1 = s_2 = s_3 = s_4 = \frac{1}{2}\mathbf{1}_\ell,\; s_5 = s_6 = \kappa\mathbf{1}_n,$$

and

$$\begin{aligned} q_1 = \mathbf{1}_m - (\varepsilon\mathbf{1}_m + R^{-1/2}z)_-,\; q_2 = \mathbf{1}_m + (\varepsilon\mathbf{1}_m + R^{-1/2}z)_+, \\ q_3 = \mathbf{1}_m - (\varepsilon\mathbf{1}_m - R^{-1/2}z)_-,\; q_4 = \mathbf{1}_m + (\varepsilon\mathbf{1}_m - R^{-1/2}z)_+, \\ q_5 = \mathbf{1}_n - (Q^{-1/2}\mu)_-,\; q_6 = \mathbf{1}_n + (Q^{-1/2}\mu)_+ \;. \end{aligned}$$

Then $(s,q,u,y) \in \mathscr{F}_+$.

5. Kalman Smoothing with PLQ penalties

Consider now a dynamic scenario, where the system state x_k evolves according to the following stochastic discrete-time linear model

$$\begin{aligned} & x_1 = x_0 + w_1 \\ & x_k = G_k x_{k-1} + w_k, \qquad k = 2,3,\ldots,N \\ & z_k = H_k x_k + v_k, \qquad k = 1,2,\ldots,N \end{aligned} \tag{5.1}$$

where x_0 is known, z_k is the m-dimensional subvector of z containing the noisy output samples collected at instant k, G_k and H_k are known matrices. Further, we consider the general case where $\{w_k\}$ and $\{v_k\}$ are mutually independent zero-mean random variables which can come from any of the densities introduced in the previous section, with positive definite covariance matrices denoted by $\{Q_k\}$ and $\{R_k\}$, respectively.

In order to formulate the Kalman smoothing problem over the entire sequence $\{x_k\}$, define

$$\begin{aligned} x &= \mathrm{vec}\{x_1, \cdots, x_N\}\,, & w &= \mathrm{vec}\{w_1, \cdots, w_N\} \\ v &= \mathrm{vec}\{v_1, \cdots, v_N\}\,, & Q &= \mathrm{diag}\{Q_1, \cdots, Q_N\} \\ R &= \mathrm{diag}\{R_1, \cdots, R_N\}\,, & H &= \mathrm{diag}\{H_1, \cdots, H_N\}, \end{aligned}$$

and

$$G = \begin{bmatrix} \mathrm{I} & 0 & & \\ -G_2 & \mathrm{I} & \ddots & \\ & \ddots & \ddots & 0 \\ & & -G_N & \mathrm{I} \end{bmatrix}$$

Then model (5.1) can be written in the form of (1.1)-(1.2), i.e.,

$$\begin{aligned} \mu &= Gx + w \\ z &= Hx + v \, , \end{aligned} \tag{5.2}$$

where $x \in \mathbb{R}^{nN}$ is the entire state sequence of interest, w is corresponding process noise, z is the vector of all measurements, v is the measurement noise, and μ is a vector of size nN with the first n-block equal to x_0, the initial state estimate, and the other blocks set to 0. This is precisely the problem (1.1)-(1.2) that began our study. The problem (1.3) becomes the classical Kalman smoothing problem with quadratic penalties. When the penalties arise from general PLQ densities, the general Kalman smoothing problem takes the form (4.3), studied in the previous section. The details are provided in the following remark.

Remark 13 *Suppose that the noises w and v in the model* (5.2) *are PLQ densities with means* 0*, variances Q and R (see Def. 9). Then, for suitable U_w, M_w, b_w, B_w and U_v, M_v, b_v, B_v and corresponding ρ_w and ρ_v we have*

$$\begin{aligned} \mathbf{p}(w) &\propto \exp\left[-\rho\left(U_w, M_w, b_w, B_w; Q^{-1/2}w\right)\right] \\ \mathbf{p}(v) &\propto \exp\left[-\rho(U_v, M_v, b_v, B_v; R^{-1/2}v)\right] \end{aligned} \tag{5.3}$$

while the MAP estimator of x in the model (5.2) *is*

$$\operatorname*{argmin}_{x\in\mathbb{R}^{nN}} \left\{ \begin{array}{l} \rho\left[U_w, M_w, b_w, B_w; Q^{-1/2}(Gx-\mu)\right] \\ +\rho\left[U_v, M_v, b_v, B_v; R^{-1/2}(Hx-z)\right] \end{array} \right\} \tag{5.4}$$

If U_w and U_v are given as in (4.9), then the system (4.6) decomposes as

$$\begin{array}{rcl} 0 &=& A_w^{\mathrm{T}}u_w + s_w - a_w \, ; \qquad 0 = A_v^{\mathrm{T}}u_v + s_v - a_v \\ 0 &=& s_w^{\mathrm{T}}q_w \, ; \qquad\qquad\quad\;\; 0 = s_v^{\mathrm{T}}q_v \\ 0 &=& \tilde{b}_w + B_w Q^{-1/2}Gd - M_w u_w - A_w q_w \\ 0 &=& \tilde{b}_v - B_v R^{-1/2}Hd - M_v u_v - A_v q_v \\ 0 &=& G^{\mathrm{T}}Q^{-\mathrm{T}/2}B_w^{\mathrm{T}}u_w - H^{\mathrm{T}}R^{-\mathrm{T}/2}B_v^{\mathrm{T}}u_v \\ 0 &\le& s_w, s_v, q_w, q_v . \end{array} \tag{5.5}$$

See the Appendix for details on deriving the KKT system. By further exploiting the decomposition shown in (5.1), we obtain the following theorem.

Theorem 14 *(PLQ Kalman Smoother Theorem) Suppose that all w_k and v_k in the Kalman smoothing model (5.1) come from PLQ densities that satisfy*

$$\mathrm{null}(M_k^w) \cap \mathrm{null}((A_k^w)^{\mathrm{T}}) = \{0\} \, , \mathrm{null}(M_k^v) \cap \mathrm{null}((A_k^v)^{\mathrm{T}}) = \{0\} \, , \forall k \, . \tag{5.6}$$

i.e. their corresponding penalties are finite-valued. Suppose further that the corresponding set $\mathscr{F}_+$ from Theorem 11 is nonempty. Then (5.4) *can be solved using an IP method, with computational complexity $O[N(n^3+m^3+l)]$, where l is the largest column dimension of the matrices $\{A_k^v\}$ and $\{A_k^w\}$.*

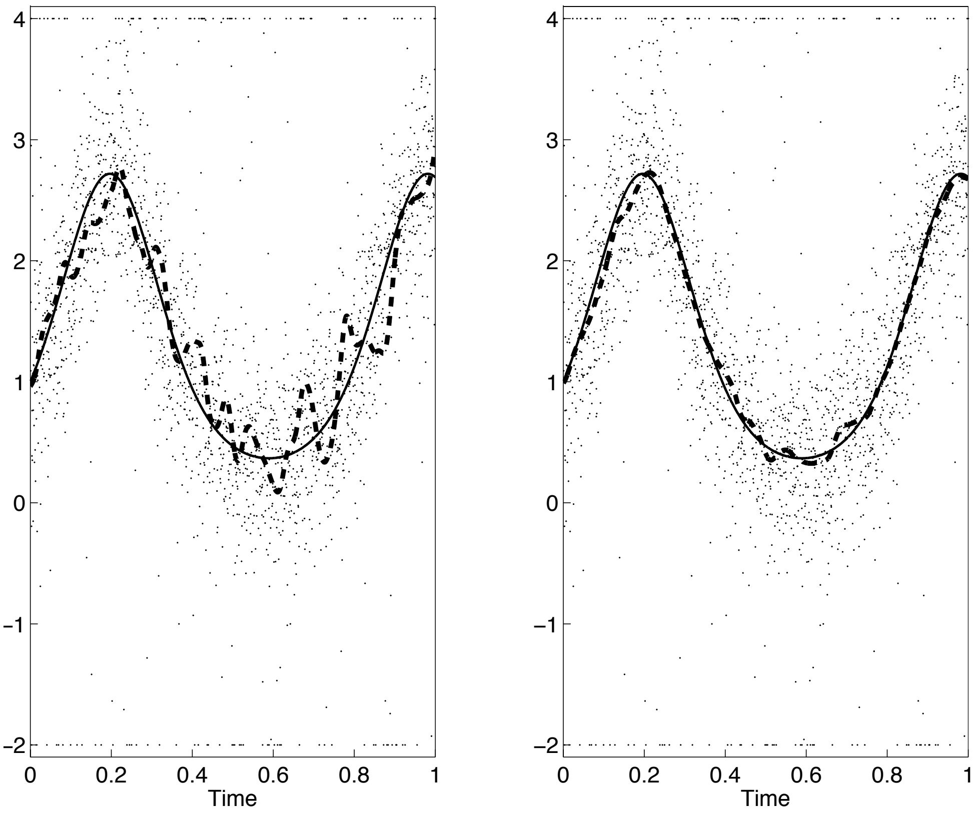

Figure 2: Simulation: measurements $(\cdot)$ with outliers plotted on axis limits (4 and -2), true function (continuous line), smoothed estimate using either the quadratic loss (dashed line, left panel) or the Vapnik's ε-insensitive loss (dashed line, right panel)

Note that the first part of this theorem, the solvability of the problem using IP methods, already follows from Theorem 11. The main contribution of the result in the dynamical system context is the computational complexity. The proof is presented in the Appendix and shows that IP methods for solving (5.4) preserve the key block tridiagonal structure of the standard smoother. If the number of IP iterations is fixed ($10-20$ are typically used in practice), general smoothing estimates can thus be computed in $O[N(n^3+m^3+l)]$ time. Notice also that the number of required operations scales favorably with l, which represents the complexity of the PLQ density encoding.

6. Numerical example

In this section we use a simulated example to test the computational scheme described in the previous section. We consider the following function

$$f(t) = \exp\left[\sin(8t)\right]$$

taken from (Dinuzzo et al., 2007). Our aim is to reconstruct f starting from 2000 noisy samples collected uniformly over the unit interval. The measurement noise v_k was generated using a mixture of two normals with $p = 0.1$ denoting the fraction from each normal; i.e.,

$$v_k \sim (1-p)\mathbf{N}(0, 0.25) + p\mathbf{N}(0, 25),$$

where $\mathbf{N}$ refers to the Normal distribution. Data are displayed as dots in Fig. 2. Note that the purpose of the second component of the normal mixture is to simulate outliers in the output data and that all the measurements exceeding vertical axis limits are plotted on upper and lower axis limits (4 and -2) to improve readability.

The initial condition $f(0) = 1$ is assumed to be known, while the difference of the unknown function from the initial condition (i.e. $f(\cdot) - 1$) is modeled as a Gaussian process given by an integrated Wiener process. This model captures the Bayesian interpretation of cubic smoothing splines (Wahba, 1990), and admits a 2-dimensional state space representation where the first component of $x(t)$, which models $f(\cdot) - 1$, corresponds to the integral of the second state component, modelled as Brownian motion. To be more specific, letting $\Delta t = 1/2000$, the sampled version of the state space model (see (Jazwinski, 1970; Oksendal, 2005) for details) is defined by

$$\begin{aligned} G_k &= \begin{bmatrix} 1 & 0 \\ \Delta t & 1 \end{bmatrix}, \qquad k = 2, 3, \ldots, 2000 \\ H_k &= \begin{bmatrix} 0 & 1 \end{bmatrix}, \qquad k = 1, 2, \ldots, 2000 \end{aligned}$$

with the autocovariance of w_k given by

$$Q_k = \lambda^2 \begin{bmatrix} \Delta t & \frac{\Delta t^2}{2} \\ \frac{\Delta t^2}{2} & \frac{\Delta t^3}{3} \end{bmatrix}, \qquad k = 1, 2, \ldots, 2000\,,$$

where λ^2 is an unknown scale factor to be estimated from the data.

The performance of two different Kalman smoothers are compared. The first (classical) estimator uses a quadratic loss function to describe the negative log of the measurement noise density and contains only λ^2 as unknown parameter. The second estimator is a Vapnik smoother relying on the ε-insensitive loss, and so depends on two unknown parameters λ^2 and ε. In both of the cases, the unknown parameters are estimated by means of a cross validation strategy where the 2000 measurements are randomly split into a training and a validation set of 1300 and 700 data points, respectively. The Vapnik smoother was implemented by exploiting the efficient computational strategy described in the previous section, see (Aravkin et al., 2011b) for specific implementation details. In this way, for each value of λ^2 and ε contained in a 10×20 grid on $[0.01, 10000] \times [0, 1]$, with λ^2 logarithmically spaced, the function estimate was rapidly obtained by the new smoother applied to the training set. Then, the relative average prediction error on the validation set was computed, see Fig. 3. The parameters leading to the best prediction were $\lambda^2 = 2.15 \times 10^3$ and $\varepsilon = 0.45$, which give a sparse solution defined by fewer than 400 support vectors. The value of λ^2 for the classical Kalman smoother was then estimated following the same strategy described above. In contrast to the Vapnik penalty, the quadratic loss does not induce any sparsity, so that, in this case, the number of support vectors equals the size of the training set.

The left and right panels of Fig. 2 display the function estimate obtained using the quadratic and the Vapnik losses, respectively. It is clear that the Gaussian estimate is heavily affected by the outliers. In contrast, as expected, the estimate coming from the Vapnik based smoother performs well over the entire time period, and is virtually unaffected by the presence of large outliers.

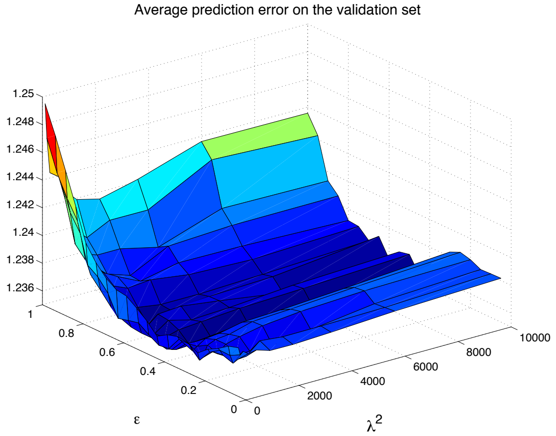

Figure 3: Estimation of the smoothing filter parameters using the Vapnik loss. Average prediction error on the validation data set as a function of the variance process λ^2 and ε.

7. Conclusions

We have presented a new theory for robust and sparse estimation using nonsmooth PLQ penalties. Using their dual representation, we first established conditions allowing the interpretation of PLQ penalties as negative logs of true probability densities, thus establishing a statistical modelling framework. In this regard, the coercivity condition characterized in Th. 7 played a central role. This condition, necessary for the statistical interpretation, underscores the importance of an idea already useful in machine learning. Specifically, coercivity of the objective (1.4) is a fundamental prerequisite in sparse and robust estimation, as it precludes directions for which the sum of the loss and the regularizer are insensitive to large parameter changes. Thus, the condition for a PLQ penalty to be a negative log of a true density also ensures that the problem is well posed in the machine learning context, i.e. the learning machine has enough control over model complexity.

In the second part of the paper, we presented a broad computational approach to solving estimation problems (1.4) using interior point methods. In the process, we derived additional conditions that guarantee the successful implementation of IP methods to compute the estimator (1.4) when x and v come from PLQ densities, and provided a theorem characterizing the convergence of IP methods for this class. The key condition required for the successful execution of IP iterations was a requirement on PLQ penalties to be finite valued, which implies non-degeneracy of the corresponding statistical distribution (the support cannot be contained in a lower-dimensional subspace). The statistical interpretation is thus strongly linked to the computational procedure.

We then applied both the statistical framework and the computational approach to the broad class of state estimation in discrete-time dynamic systems, extending the classical formulations to allow dynamics and measurement noise to come from any PLQ densities. Moreover, we showed that the classical computational efficiency results can be preserved when the general IP approach is applied in the state estimation context; specifically, PLQ Kalman smoothing can always be performed with a number of operations that is linear in the length of the time series, as in the quadratic case. The computational framework presented therefore allows the broad application of interior point methods to a wide class of smoothing problems of interest to practitioners. The powerful algorithmic scheme designed here, together with the breadth and significance of the new statistical framework presented, underscores the practical utility and flexibility of this approach. We believe that this perspective on modeling, robust/sparse estimation and Kalman smoothing will be useful in a number of applications in the years ahead.

8. Appendix

8.1 Preliminaries

We begin by supplementing subsection 2.1 with the definition and characterization of *lineality*, a concept which will be useful in the rest of the Appendix.

Definition 15 *(Lineality). Define the lineality of convex cone K, denoted* $\mathrm{lin}(K)$*, to be $K \cap -K$. Since K is a convex cone,* $\mathrm{lin}(K)$ *is the largest subspace contained in K.*

Lemma 16 *(Characterization of lineality, (Rockafellar, 1970, Theorem 14.6)). Let K be any closed set containing the origin. Then* $\mathrm{lin}(K) = [K^\circ]^\perp$.

Corollary 17 *(Characterization of* $\mathrm{aff}\, K^\circ$*) Taking the perp of the characterization in Lemma 16, the affine hull of the polar of a closed convex cone K is given by* $\mathrm{aff}\, K^\circ = \mathrm{lin}(K)^\perp$.

8.2 Proof of Theorem 3

Lemma 18 *(Polars, linear transformations, and shifts) Let $K \subset \mathbb{R}^n$ be a closed convex cone, $b \in K^\circ$, and $B \in \mathbb{R}^{n\times k}$. Then we have*

$$(B^{\mathrm{T}}K)^\circ \subset B^{-1}[K^\circ - b] \; .$$

Proof Recall that a convex cone is closed under addition. Then for any $b \in K^\circ$, we have $K^\circ + b \subset K^\circ$, and hence $K^\circ \subset K^\circ - b$. By (Rockafellar, 1970, Corollary 16.3.2) we get

$$(B^{\mathrm{T}}K)^\circ = B^{-1}K^\circ \subset B^{-1}[K^\circ - b] \; .$$

■

Corollary 19 *Let K be a closed convex cone, and $B \in \mathbb{R}^{n\times k}$. If $b \in K^\circ$, then*

$$\left(B^{\mathrm{T}}[\mathrm{lin}(K)]\right)^\perp \subset \mathrm{aff}(B^{-1}[K^\circ - b]).$$

Proof By Lemma 18,

$$\mathrm{aff}(B^{-1}[K^\circ - b]) \supset \mathrm{aff}[(B^{\mathrm{T}}K)^\circ] = \mathrm{lin}[B^{\mathrm{T}}K]^\perp \; ,$$

where the last equality is by Corollary 17. Since B^{T} is a linear transformation, we have $\mathrm{lin}(B^{\mathrm{T}}K) = B^{\mathrm{T}}\mathrm{lin}(K)$. ■

Lemma 20 *Let $K \subset \mathbb{R}^n$ be a closed convex cone, $b \in \text{aff}[K^\circ]$, and $B \in \mathbb{R}^{n\times k}$. Then*

$$\text{aff}(B^{-1}[K^\circ - b]) \subset B^{-1}[\text{lin}(K)]^\perp \subset B^{-1}\text{aff}[K^\circ - b] \ .$$

Proof If $w \in \text{aff}\left(B^{-1}[K^\circ - b]\right)$, for some finite N we can find sets $\{\lambda_i\} \subset \mathbb{R}$ and $\{w_i\} \subset B^{-1}[K^\circ - b]$ such that $\sum_{i=1}^N \lambda_i = 1$ and $\sum_{i=1}^N \lambda_i w_i = w$. For each w_i, we have $Bw_i \in K^\circ - b$, so $b + Bw_i \in K^\circ$. Then

$$b + Bw = \sum_{i=1}^N \lambda_i (b + Bw_i) \in \text{aff}[K^\circ] = \text{lin}(K)^\perp.$$

Since $b \in \text{lin}(K)^\perp$ by assumption, we have $Bw \in \text{lin}(K)^\perp$, and so $w \in B^{-1}[\text{lin}(K)^\perp]$.

Next, starting with $w \in B^{-1}[\text{lin}(K)^\perp]$ we have $Bw \in \text{lin}(K)^\perp$ and so $b + Bw \in \text{lin}(K)^\perp$ since $\text{lin}(K)^\perp$ is a subspace and $b \in \text{lin}(K)^\perp$. Then for some finite $\tilde{N}$ we can find sets $\{\lambda_i\} \subset \mathbb{R}$ and $\{v_i\} \subset K^\circ$ such that $\sum_{i=1}^{\tilde{N}} \lambda_i = 1$ and $\sum_{i=1}^{\tilde{N}} \lambda_i v_i = b + Bw$. Subtracting b from both sides, we have $\sum_{i=1}^{\tilde{N}} \lambda_i (v_i - b) = Bw$, so in particular $Bw \in \text{aff}[K^\circ - b]$. Then $w \in B^{-1}\text{aff}[K^\circ - b]$. ■

Theorem 21 *Let $K \subset \mathbb{R}^n$ be a closed convex cone, $b \in \mathbb{R}^n$, and $B \in \mathbb{R}^{n\times k}$. If $b \in K^\circ$, then*

$$[B^{\mathrm{T}}\text{lin}(K)]^\perp = \text{aff}\left(B^{-1}[K^\circ - b]\right) = B^{-1}[\text{lin}(K)^\perp]$$

.

Proof From Corollary 19 and Lemma 20, we immediately have

$$[B^{\mathrm{T}}\text{lin}(K)]^\perp \subset \text{aff}\left(B^{-1}[K^\circ - b]\right) \subset B^{-1}[\text{lin}(K)^\perp].$$

Note that for any subspace C, $C^\perp = C^\circ$. Then by (Rockafellar, 1970, Corollary 16.3.2), $[B^{\mathrm{T}}\text{lin}(K)]^\perp = B^{-1}[\text{lin}(K)^\perp]$. ■

The proof of Theorem 3 now follows from Lemma 18 and Theorem 21.

8.3 Proof of Theorem 6

Using the characterization of a piecewise quadratic function from (Rockafellar and Wets, 1998, Definition 10.20), the effective domain of $\rho(y)$ can be represented as the union of finitely many polyhedral sets U_i, relative to each of which $\rho(y)$ is given by an expression of the form $\frac{1}{2}\langle y, A_i y\rangle + \langle a_i, y\rangle + \alpha_i$ for some scalar $\alpha_i \in \mathbb{R}$, vector $a_i \in \mathbb{R}^n$ and symmetric positive semidefinite matrix $A_i \in \mathbb{R}^{n\times n}$. Since $\rho(y)$ is coercive, we claim that on each unbounded U_i there must be some constants P_i and $\beta_i > 0$ so that for $\|y\| \geq P_i$ we have $\rho(y) \geq \beta_i \|y\|$. Otherwise, we can find an index set J such that $\rho(y_j) \leq \beta_j \|y_j\|$, where $\beta_j \downarrow 0$ and $\|y_j\| \uparrow \infty$. Without loss of generality, suppose $\frac{y_j}{\|y_j\|}$converges to $\bar{y} \in U_i^\infty$, by (Rockafellar, 1970, Theorem 8.2). By assumption, $\frac{\rho(y_j)}{\|y_j\|} \downarrow 0$, and we have

$$\frac{\rho(y_j)}{\|y_j\|} = \|y_j\| \left\langle \frac{y_j}{\|y_j\|}, A_i \frac{y_j}{\|y_j\|} \right\rangle + \left\langle a_i, \frac{y_j}{\|y_j\|} \right\rangle + \frac{\alpha_i}{\|y_j\|}.$$

Taking the limit of both sides over J we see that $\|y_j\| \left\langle \frac{y_j}{\|y_j\|}, A_i \frac{y_j}{\|y_j\|} \right\rangle$ must converge to a finite value. But this is only possible if $\langle \bar{y}, A_i \bar{y}\rangle = 0$, so in particular we must have $\bar{y} \in \text{null}(A_i)$. Note also that

$\langle a_i, \bar{y}\rangle \le 0$, by taking the limit over J of

$$\frac{\rho(y_j)}{\|y_j\|} \ge \left\langle a_i, \frac{y_j}{\|y_j\|}\right\rangle + \frac{\alpha}{\|y_i\|},$$

so for any $x_0 \in U_i$ and $\lambda > 0$ we have $x_0 + \lambda\bar{y} \in U_i$ since $\bar{y} \in U_i^\infty$ and

$$\rho(x_0 + \lambda\bar{y}) \le \rho(x_0) + \alpha_i,$$

so in particular ρ stays bounded as $\lambda \uparrow \infty$ and cannot be coercive.

The integrability of the (nonnegative) function $\exp[-\rho(y)]$ is now clear. Recall that the effective domain of ρ can be represented as the union of finitely many polyhedral sets U_i, and for each unbounded such U_i we have shown $\exp[-\rho(y)] \le \exp[-\beta_i\|y\|]$ off of some bounded subset of U_i. The result now follows from the bounded convergence theorem.

8.4 Proof of Theorem 7

First observe that $B^{-1}[\text{cone}(U)]^\circ = [B^\text{T}\text{cone}(U)]^\circ$ by (Rockafellar, 1970, Corollary 16.3.2).

Suppose that $\hat{y} \in B^{-1}[\text{cone}(U)]^\circ$, and $\hat{y} \neq 0$. Then $B\hat{y} \in \text{cone}(U)$, and $B\hat{y} \neq 0$ since B is injective, and we have

$$\begin{array}{rcl} \rho(t\hat{y}) & = & \sup_{u\in U}\langle b + tB\hat{y}, u\rangle - \frac{1}{2}u^\text{T}Mu \\ & = & \sup_{u\in U}\langle b, u\rangle - \frac{1}{2}u^\text{T}Mu + t\langle B\hat{y}, u\rangle \\ & \le & \sup_{u\in U}\langle b, u\rangle - \frac{1}{2}u^\text{T}Mu \\ & \le & \rho(U, M, 0, I; b), \end{array}$$

so $\rho(t\hat{y})$ stays bounded even as $t \to \infty$, and so ρ cannot be coercive.

Conversely, suppose that ρ is not coercive. Then we can find a sequence $\{y_k\}$ with $\|y_k\| > k$ and a constant P so that $\rho(y_k) \le P$ for all $k > 0$. Without loss of generality, we may assume that $\frac{y_k}{\|y_k\|} \to \bar{y}$.

Then by definition of ρ, we have for all $u \in U$

$$\begin{array}{l} \langle b + By_k, u\rangle - \frac{1}{2}u^\text{T}Mu \le P \\ \langle b + By_k, u\rangle \le K + \frac{1}{2}u^\text{T}Mu \\ \langle \frac{b+By_k}{\|y_k\|}, u\rangle \le \frac{K}{\|y_k\|} + \frac{1}{2\|y_k\|}u^\text{T}Mu \end{array}$$

Note that $\bar{y} \neq 0$, so $B\bar{y} \neq 0$. When we take the limit as $k \to \infty$, we get $\langle B\bar{y}, u\rangle \le 0$. From this inequality we see that $B\bar{y} \in [\text{cone}(U)]^\circ$, and so $\bar{y} \in B^{-1}[\text{cone}(U)]^\circ$.

8.5 Proof of Theorem 11

Proof (i) Using standard elementary row operations, reduce the matrix

$$F_\gamma^{(1)} := \begin{bmatrix} I & 0 & A^\text{T} & 0 \\ D(q) & D(s) & 0 & 0 \\ 0 & -A & -M & B \\ 0 & 0 & B^\text{T} & 0 \end{bmatrix} \tag{8.1}$$

to

$$\begin{bmatrix} I & 0 & A^{\mathrm{T}} & 0 \\ 0 & D(s) & -D(q)A^{\mathrm{T}} & 0 \\ 0 & 0 & -T & B \\ 0 & 0 & B^{\mathrm{T}} & 0 \end{bmatrix},$$

where $T = M + AD(q)D(s)^{-1}A^{\mathrm{T}}$. The matrix T is invertible since $\mathrm{null}(M) \cap \mathrm{null}(C^{\mathrm{T}}) = \{0\}$. Hence, we can further reduce this matrix to the block upper triangular form

$$\begin{bmatrix} I & 0 & A^{\mathrm{T}} & 0 \\ 0 & D(s) & -D(q)C^{\mathrm{T}} & 0 \\ 0 & 0 & -T & B \\ 0 & 0 & 0 & -B^{\mathrm{T}}T^{-1}B \end{bmatrix}.$$

Since B is injective, the matrix $B^{\mathrm{T}}T^{-1}B$ is also invertible. Hence this final block upper triangular is invertible proving Part (i).

(ii) Let $(s,q) \in \widehat{\mathscr{F}_+}$ and choose (u_i, y_i) so that $(s,q,u_i,y_i) \in \mathscr{F}_+$ for $i = 1,2$. Set $u := u_1 - u_2$ and $y := y_1 - y_2$. Then, by definition,

$$0 = A^{\mathrm{T}}u, \; 0 = By - Mu, \text{ and } 0 = B^{\mathrm{T}}u\,. \tag{8.2}$$

Multiplying the second of these equations on the left by u and utilizing the third as well as the positive semi-definiteness of M, we find that $Mu = 0$. Hence, $u \in \mathrm{null}(M) \cap \mathrm{null}(A^{\mathrm{T}}) = \{0\}$, and so $By = 0$. But then $y = 0$ as B is injective.

(iii) Let $(\hat{s}, \hat{q}, \hat{u}, \hat{y}) \in \mathscr{F}_+$ and $(s,q,u,y) \in \mathscr{F}_+(\tau)$. Then, by (4.6),

$$\begin{aligned} (s-\hat{s})^{\mathrm{T}}(q-\hat{q}) &= [(a - A^{\mathrm{T}}u) - (a - A^{\mathrm{T}}\hat{u})]^{\mathrm{T}}(q-\hat{q}) \\ &= (\hat{u}-u)^{\mathrm{T}}(Aq - A\hat{q}) \\ &= (\hat{u}-u)^{\mathrm{T}}[(b + By - Mu) - (b + B\hat{b} - M\hat{u})] \\ &= (\hat{u}-u)^{\mathrm{T}}M(\hat{u}-u) \\ &\geq 0. \end{aligned}$$

Hence,

$$\tau + \hat{s}^{\mathrm{T}}\hat{q} \geq s^{\mathrm{T}}y + \hat{s}^{\mathrm{T}}\hat{q} \geq s^{\mathrm{T}}\hat{y} + y^{\mathrm{T}}\hat{s} \geq \xi\, \|(s,q)\|_1,$$

where $\xi = \min\{\hat{s}_i,\ \hat{q}_i \mid i = 1, \ldots, \ell\} > 0$. Therefore, the set

$$\widehat{\mathscr{F}_+}(\tau) = \{(s,q) \mid (s,q,u,y) \in \mathscr{F}_+(\tau)\}$$

is bounded. Now suppose the set $\mathscr{F}_+(\tau)$ is not bounded. Then there exits a sequence $\{(s_\nu, q_\nu, u_\nu, y_\nu)\} \subset \mathscr{F}_+(\tau)$ such that $\|(s_\nu, q_\nu, u_\nu, y_\nu)\| \uparrow +\infty$. Since $\widehat{\mathscr{F}_+}(\tau)$ is bounded, we can assume that $\|(u_\nu, y_\nu)\| \uparrow +\infty$ while $\|(s_\nu, q_\nu)\|$ remains bounded. With no loss in generality, we may assume that there exits $(u,y) \neq (0,0)$ such that $(u_\nu, y_\nu)/\|(u_\nu, y_\nu)\| \to (u,y)$. By dividing (4.6) by $\|(u_\nu, y_\nu)\|$ and taking the limit, we find that (8.2) holds. But then, as in (8.2), $(u,y) = (0,0)$. This contradiction yields the result.

(iv) We first show existence. This follows from a standard continuation argument. Let $(\hat{s}, \hat{q}, \hat{u}, \hat{y}) \in \mathscr{F}_+$ and $v \in \mathbb{R}^{\ell}_{++}$. Define

$$F(s,q,u,y,t) = \begin{bmatrix} s + A^{\mathrm{T}}u - a \\ D(q)D(s)\mathbf{1} - [(1-t)\hat{v} + tv] \\ By - Mu - Aq \\ B^{\mathrm{T}}u + b \end{bmatrix}, \tag{8.3}$$

where $\hat{g} := (\hat{s}_1\hat{y}_1, \ldots, \hat{s}_\ell\hat{y}_\ell)^{\mathrm{T}}$. Note that

$$F(\hat{s}, \hat{q}, \hat{u}, \hat{y}, 0) = 0 \text{ and, by Part (i), } \nabla_{(s,q,u,y)}F(\hat{s}, \hat{q}, \hat{u}, \hat{y}, 0)^{-1} \text{ exists.}$$

The Implicit Function Theorem implies that there is a $\tilde{t} > 0$ and a differentiable mapping $t \mapsto (s(t), q(t), u(t), y(t))$ on $[0, \tilde{t})$ such that

$$F[s(t), q(t), u(t), y(t), t] = 0 \text{ on } [0, \tilde{t}).$$

Let $\bar{t} > 0$ be the largest such $\tilde{t}$ on $[0,1]$. Since

$$\{[s(t), q(t), u(t), y(t)] \,|\, t \in [0, \bar{t})\} \subset \mathscr{F}_+(\bar{\tau}),$$

where $\bar{\tau} = \max\{\mathbf{1}^{\mathrm{T}}\hat{g}, \mathbf{1}^{\mathrm{T}}g\}$, Part (iii) implies that there is a sequence $t_i \to \bar{t}$ and a point $(\bar{s}, \bar{q}, \bar{u}, \bar{y})$ such that $[s(t_i), q(t_i), u(t_i), y(t_i)] \to (\bar{s}, \bar{q}, \bar{u}, \bar{y})$. By continuity $F(\bar{s}, \bar{q}, \bar{u}, \bar{y}, \bar{t}) = 0$. If $\bar{t} = 1$, we are done; otherwise, apply the Implicit Function Theorem again at $(\bar{s}, \bar{q}, \bar{u}, \bar{y}, \bar{t})$ to obtain a contradiction to the maximality of $\bar{t}$.

We now show uniqueness. By Part (ii), we need only establish the uniqueness of (s,q). Let $(s^{\nu}, q^{\nu}) \in \widehat{\mathscr{F}_+}$ be such that $g = (s_{j(1)}q_{j(1)}, s_{j(2)}q_{j(2)}, \ldots, s_{j(\ell)}q_{j(\ell)})^{\mathrm{T}}$, where $s_{j(i)}$ denotes the ith element of s_j, and $j = 1,2$. As in Part (iii), we have $(s_1 - s_2)^{\mathrm{T}}(q_1 - q_2) = (u_1 - u_2)^{\mathrm{T}}M((u_1 - u_2) \geq 0$, and, for each $i = 1, \ldots, \ell$, $s_{1(i)}q_{1(i)} = s_{2(i)}q_{2(i)} = g_i > 0$. If $(s_1, q_1) \neq (s_2, q_2)$, then, for some $i \in \{1, \ldots, \ell\}$, $(s_{1(i)} - s_{2(i)})(q_{1(i)} - q_{2(i)}) \geq 0$ and either $s_{1(i)} \neq s_{2(i)}$ or $q_{1(i)} \neq q_{2(i)}$. If $s_{1(i)} > s_{2(i)}$, then $q_{1(i)} \geq q_{2(i)} > 0$ so that $g_i = s_{1(i)}q_{1(i)} > s_{2(i)}q_{2(i)} = g_i$, a contradiction. So with out loss in generality (by exchanging (s_1, q_1) with (s_2, q_2) if necessary), we must have $q_{1(i)} > q_{2(i)}$. But then $s_{1(i)} \geq s_{2(i)} > 0$, so that again $g_i = s_{1(i)}q_{1(i)} > s_{2(i)}q_{2(i)} = g_i$, and again a contradiction. Therefore, (s,q) is unique.

(v) Apply Part (iv) to get a point on the central path and then use the continuation argument to trace out the central path. The differentiability follows from the implicit function theorem.

(vi) Part (iii) allows us to apply a standard compactness argument to get the existence of cluster points and the continuity of $F_\gamma(s,q,u,y)$ in all of its arguments including γ implies that all of these cluster points solve (4.6). ■

8.6 Details for Remark 13

The Lagrangian for (5.4) for feasible (x, u_w, u_v) is

$$L(x, u_w, u_v) = \left\langle \begin{bmatrix} \tilde{b}_w \\ \tilde{b}_v \end{bmatrix}, \begin{bmatrix} u_w \\ u_v \end{bmatrix} \right\rangle - \frac{1}{2} \begin{bmatrix} u_w \\ u_v \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} M_w & 0 \\ 0 & M_v \end{bmatrix} \begin{bmatrix} u_w \\ u_v \end{bmatrix} - \left\langle \begin{bmatrix} u_w \\ u_v \end{bmatrix}, \begin{bmatrix} -B_w Q^{-1/2} G \\ B_v R^{-1/2} H \end{bmatrix} x \right\rangle \tag{8.4}$$

where $\tilde{b}_w = b_w - B_w Q^{-1/2}\tilde{x}_0$ and $\tilde{b}_v = b_v - B_v R^{-1/2} z$. The associated optimality conditions for feasible (x, u_w, u_v) are given by

$$\begin{array}{l} G^{\mathrm{T}} Q^{-\mathrm{T}/2} B_w^{\mathrm{T}} \bar{u}_w - H^{\mathrm{T}} R^{-\mathrm{T}/2} B_v^{\mathrm{T}} \bar{u}_v = 0 \\ \tilde{b}_w - M_w \bar{u}_w + B_w Q^{-1/2} G \bar{x} \in N_{U_w}(\bar{u}_w) \\ \tilde{b}_v - M_v \bar{u}_v - B_v R^{-1/2} H \bar{x} \in N_{U_v}(\bar{u}_v) \ , \end{array} \tag{8.5}$$

where $N_C(r)$ denotes the normal cone to the set C at the point r (see (Rockafellar, 1970) for details).

Since U_w and U_v are polyhedral, we can derive explicit representations of the normal cones $N_{U_w}(\bar{u}_w)$ and $N_{U_v}(\bar{u}_v)$. For a polyhedral set $U \subset \mathbb{R}^m$ and any point $\bar{u} \in U$, the normal cone $N_U(\bar{u})$ is polyhedral. Indeed, relative to any representation

$$U = \{u | A^{\mathrm{T}} u \le a\}$$

and the active index set $I(\bar{u}) := \{i | \langle A_i, \bar{u} \rangle = a_i\}$, where A_i denotes the ith column of A, we have

$$N_U(\bar{u}) = \left\{ \begin{array}{r} q_1 A_1 + \cdots + q_m A_m \mid q_i \ge 0 \text{ for } i \in I(\bar{u}) \\ q_i = 0 \text{ for } i \notin I(\bar{u}) \end{array} \right\}. \tag{8.6}$$

Using (8.6), Then we may rewrite the optimality conditions (8.5) more explicitly as

$$\begin{array}{l} G^{\mathrm{T}} Q^{-\mathrm{T}/2} B_w^{\mathrm{T}} \bar{u}_w - H^{\mathrm{T}} R^{-\mathrm{T}/2} B_v^{\mathrm{T}} \bar{u}_v = 0 \\ \tilde{b}_w - M_w \bar{u}_w + B_w Q^{-1/2} G \bar{d} = A_w q_w \\ \tilde{b}_v - M_v \bar{u}_v - B_v R^{-1/2} H \bar{d} = A_v q_v \\ \{q_v \ge 0 | q_{v(i)} = 0 \text{ for } i \notin I(\bar{u}_v)\} \\ \{q^w \ge 0 | q_{w(i)} = 0 \text{ for } i \notin I(\bar{u}_w)\} \end{array} \tag{8.7}$$

where $q_{v(i)}$ and $q_{w(i)}$ denote the ith elements of q_v and q_w. Define slack variables $s_w \ge 0$ and $s_v \ge 0$ as follows:

$$\begin{array}{rcl} s_w & = & a_w - A_w^{\mathrm{T}} u_w \\ s_v & = & a_v - A_v^{\mathrm{T}} u_v . \end{array} \tag{8.8}$$

Note that we know the entries of $q_{w(i)}$ and $q_{v(i)}$ are zero if and only if the corresponding slack variables $s_{v(i)}$ and $s_{w(i)}$ are nonzero, respectively. Then we have $q_w^{\mathrm{T}} s_w = q_v^{\mathrm{T}} s_v = 0$. These equations are known as the complementarity conditions. Together, all of these equations give system (5.5).

8.7 Proof of Theorem 14

IP methods apply a damped Newton iteration to find the solution of the relaxed KKT system $F_\gamma = 0$, where

$$F_\gamma \begin{pmatrix} s_w \\ s_v \\ q_w \\ q_v \\ u_w \\ u_v \\ x \end{pmatrix} = \begin{bmatrix} A_w^{\mathrm{T}} u_w + s_w - a_w \\ A_v^{\mathrm{T}} u_v + s_v - a_v \\ D(q_w) D(s_w) \mathbf{1} - \gamma \mathbf{1} \\ D(q_v) D(s_v) \mathbf{1} - \gamma \mathbf{1} \\ \tilde{b}_w + B_w Q^{-1/2} G d - M_w u_w - A_w q_w \\ \tilde{b}_v - B_v R^{-1/2} H d - M_v u_v - A_v q_v \\ G^{\mathrm{T}} Q^{-\mathrm{T}/2} B_w^{\mathrm{T}} u_w - H^{\mathrm{T}} R^{-\mathrm{T}/2} B_v^{\mathrm{T}} \bar{u}_v \end{bmatrix} .$$

This entails solving the system

$$F_\gamma^{(1)} \begin{pmatrix} s_w \\ s_v \\ q_w \\ q_v \\ u_w \\ u_v \\ x \end{pmatrix} \begin{bmatrix} \Delta s_w \\ \Delta s_v \\ \Delta q_w \\ \Delta q_v \\ \Delta u_w \\ \Delta u_v \\ \Delta x \end{bmatrix} = -F_\gamma \begin{pmatrix} s_w \\ s_v \\ q_w \\ q_v \\ u_w \\ u_v \\ x \end{pmatrix} , \tag{8.9}$$

where the derivative matrix $F_\gamma^{(1)}$ is given by

$$\begin{bmatrix} I & 0 & 0 & 0 & (A_w)^{\mathrm{T}} & 0 & 0 \\ 0 & I & 0 & 0 & 0 & (A_v)^{\mathrm{T}} & 0 \\ D(q_w) & 0 & D(s_w) & 0 & 0 & 0 & 0 \\ 0 & D(q_v) & 0 & D(s_v) & 0 & 0 & 0 \\ 0 & 0 & -A_w & 0 & -M_w & 0 & B_w Q^{-1/2} G \\ 0 & 0 & 0 & -A_v & 0 & -M_v & -B_v R^{-1/2} H \\ 0 & 0 & 0 & 0 & G^{\mathrm{T}} Q^{-\mathrm{T}/2} B_w^{\mathrm{T}} & -H^{\mathrm{T}} R^{-\mathrm{T}/2} B_v^{\mathrm{T}} & 0 \end{bmatrix} \tag{8.10}$$

We now show the row operations necessary to reduce the matrix $F_\gamma^{(1)}$ in (8.10) to upper block triangular form. After each operation, we show only the row that was modified.

$$\begin{aligned} &\mathrm{row}_3 \leftarrow \mathrm{row}_3 - D(q_w)\,\mathrm{row}_1 \\ &\begin{bmatrix} 0 & 0 & D(s_w) & 0 & -D(q_w)A_w^{\mathrm{T}} & 0 & 0 \end{bmatrix} \\ &\mathrm{row}_4 \leftarrow \mathrm{row}_4 - D(q_v)\,\mathrm{row}_2 \\ &\begin{bmatrix} 0 & 0 & 0 & D(s_v) & 0 & -D(q_v)A_v^{\mathrm{T}} & 0 \end{bmatrix} \\ &\mathrm{row}_5 \leftarrow \mathrm{row}_5 + A_w D(s_w)^{-1}\,\mathrm{row}_3 \\ &\begin{bmatrix} 0 & 0 & 0 & 0 & -T_w & 0 & B_w Q^{-1/2} G \end{bmatrix} \\ &\mathrm{row}_6 \leftarrow \mathrm{row}_6 + A_v D(s_v)^{-1}\,\mathrm{row}_4 \\ &\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & -T_v & -B_v R^{-1/2} H \end{bmatrix} . \end{aligned}$$

In the above expressions,

$$\begin{aligned} T_w &:= M_w + A_w D(s_w)^{-1} D(q_w) A_w^{\mathrm{T}} \\ T_v &:= M_v + A_v D(s_v)^{-1} D(q_v) A_v^{\mathrm{T}} , \end{aligned} \tag{8.11}$$

where $D(s_w)^{-1}D(q_w)$ and $D(s_v)^{-1}D(q_v)$ are always full-rank diagonal matrices, since the vectors s_w, q_w, s_v, q_v. Matrices T_w and T_v are invertible as long as the PLQ densities for w and v satisfy (4.10).

Remark 22 *(Block diagonal structure of T in i.d. case) Suppose that y is a random vector, $y = \mathrm{vec}(\{y_k\})$, where each y_i is itself a random vector in $\mathbb{R}^{m(i)}$, from some PLQ density*
$\mathbf{p}(y_i) \propto \exp[-c_2\rho(U_i, M_i, 0, I; \cdot)]$, and all y_i are independent. Let $U_i = \{u : A_i^{\mathrm{T}} u \le a_i\}$. Then the matrix T_ρ is given by $T_\rho = M + ADA^{\mathrm{T}}$ where $M = \mathrm{diag}[M_1, \cdots, M_N]$, $A = \mathrm{diag}[A_1, \cdots, A_N]$, $D = \mathrm{diag}[D_1, \cdots, D_N]$, and $\{D_i\}$ are diagonal with positive entries. Moreover, T_ρ is block diagonal, with ith diagonal block given by $M_i + A_i D_i A_i^{\mathrm{T}}$.

From Remark 22, the matrices T_w and T_v in (8.11) are block diagonal provided that $\{w_k\}$ and $\{v_k\}$ are independent vectors from any PLQ densities.

We now finish the reduction of $F_\gamma^{(1)}$ to upper block triangular form:

$$\text{row}_7 \leftarrow \text{row}_7 + \left(G^{\mathrm{T}} Q^{-\mathrm{T}/2} B_w^{\mathrm{T}} T_w^{-1}\right)\text{row}_5 - \left(H^{\mathrm{T}} R^{-\mathrm{T}/2} B_v^{\mathrm{T}} T_v^{-1}\right)\text{row}_6$$

$$\begin{bmatrix}
I & 0 & 0 & 0 & (A_w)^{\mathrm{T}} & 0 & 0 \\
0 & I & 0 & 0 & 0 & (A_v)^{\mathrm{T}} & 0 \\
0 & 0 & S_w & 0 & -Q_w(A_w)^{\mathrm{T}} & 0 & 0 \\
0 & 0 & 0 & S_v & 0 & -Q_v(A_v)^{\mathrm{T}} & 0 \\
0 & 0 & 0 & 0 & -T_w & 0 & B_w Q^{-1/2} G \\
0 & 0 & 0 & 0 & 0 & -T_v & -B_v R^{-1/2} H \\
0 & 0 & 0 & 0 & 0 & 0 & \Omega
\end{bmatrix}$$

where

$$\Omega = \Omega_G + \Omega_H = G^{\mathrm{T}} Q^{-\mathrm{T}/2} B_w^{\mathrm{T}} T_w^{-1} B_w Q^{-1/2} G + H^{\mathrm{T}} R^{-\mathrm{T}/2} B_v^{\mathrm{T}} T_v^{-1} B_v R^{-1/2} H. \tag{8.12}$$

Note that Ω is symmetric positive definite. Note also that Ω is block tridiagonal, since

1. Ω_H is block diagonal.
2. $Q^{-\mathrm{T}/2} B_w^{\mathrm{T}} T_w^{-1} B_w Q^{-1/2}$ is block diagonal, and G is block bidiagonal, hence Ω_G is block tridiagonal.

Solving system (8.9) requires inverting the block diagonal matrices T_v and T_w at each iteration of the damped Newton's method, as well as solving an equation of the form $\Omega \Delta x = \rho$. The matrices T_v and T_w are block diagonal, with sizes Nn and Nm, assuming m measurements at each time point. Given that they are invertible (see (4.10)), these inversions take $O(Nn^3)$ and $O(Nm^3)$ time. Since Ω is block tridiagonal, symmetric, and positive definite, $\Omega \Delta x = \rho$ can be solved in $O(Nn^3)$ time using the block tridiagonal algorithm in (Bell, 2000). The remaining four back solves required to solve (8.9) can each be done in $O(Nl)$ time, where we assume that $A_{v(k)} \in \mathbb{R}^{n \times l}$ and $A_{w(k)} \in \mathbb{R}^{m \times l}$ at each time point k.

References

B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, N.J., USA, 1979.

A.Y. Aravkin. *Robust Methods with Applications to Kalman Smoothing and Bundle Adjustment*. PhD thesis, University of Washington, Seattle, WA, June 2010.

A.Y. Aravkin, B.M. Bell, J.V. Burke, and G. Pillonetto. An ℓ_1-Laplace robust Kalman smoother. *IEEE Transactions on Automatic Control*, 2011a.

A.Y. Aravkin, B.M. Bell, J.V. Burke, and G. Pillonetto. Learning using state space kernel machines. In *Proc. IFAC World Congress 2011*, Milan, Italy, 2011b.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

B.M. Bell. The marginal likelihood for parameters in a discrete Gauss-Markov process. *IEEE Transactions on Signal Processing*, 48(3):626–636, August 2000.

R. Brockett. *Finite Dimensional Linear Systems*. John Wiley and Sons, Inc., 1970.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39:1–49, 2001.

F. Dinuzzo. Analysis of fixed-point and coordinate descent algorithms for regularized kernel methods. *IEEE Transactions on Neural Networks*, 22(10):1576 –1587, 2011.

F. Dinuzzo, M. Neve, G. De Nicolao, and U. P. Gianazza. On the representer theorem and equivalent degrees of freedom of SVR. *Journal of Machine Learning Research*, 8:2467–2495, 2007.

D. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, 2006.

B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–150, 2000.

S. Farahmand, G.B. Giannakis, and D. Angelosante. Doubly robust smoothing of dynamical processes via outlier sparsity constraints. *IEEE Transactions on Signal Processing*, 59:4529–4543, 2011.

M.C. Ferris and T.S. Munson. Interior-point methods for massive support vector machines. *SIAM Journal on Optimization*, 13(3):783 – 804, 2003.

S. Fine and K. Scheinberg. Efficient svm training using low-rank kernel representations. *J. Mach. Learn. Res.*, 2:243 –264, 2001.

J. Gao. Robust l1 principal component analysis and its Bayesian variational inference. *Neural Computation*, 20(2):555–572, February 2008.

A. Gelb. *Applied Optimal Estimation*. The M.I.T. Press, Cambridge, MA, 1974.

T. J. Hastie and R. J. Tibshirani. Generalized additive models. In *Monographs on Statistics and Applied Probability*, volume 43. Chapman and Hall, London, UK, 1990.

T. J. Hastie, R. J. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer, Canada, 2001.

P.J. Huber. *Robust Statistics*. Wiley, 1981.

A. Jazwinski. *Stochastic Processes and Filtering Theory*. Dover Publications, Inc, 1970.

T. Joachims, editor. *Making large-scale support vector machine learning practical*. MIT Press, Cambridge, MA, USA, 1998.

S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale ℓ_1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):606 – 617, 2007.

M. Kojima, N. Megiddo, T. Noma, and A. Yoshise. *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, volume 538 of *Lecture Notes in Computer Science*. Springer Verlag, Berlin, Germany, 1991.

C.J. Lin. On the convergence of the decomposition method for support vector machines. *IEEE Transactions on Neural Networks*, 12(12):1288 –1298, 2001.

S. Lucidi, L. Palagi, A. Risi, and M. Sciandrone. A convergent decomposition algorithm for support vector machines. *Comput. Optim. Appl.*, 38(2):217 –234, 2007.

D.J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

D.J.C. Mackay. Bayesian non-linear modelling for the prediction competition. *ASHRAE Trans.*, 100(2):3704–3716, 1994.

A. Nemirovskii and Y. Nesterov. *Interior-Point Polynomial Algorithms in Convex Programming*, volume 13 of *Studies in Applied Mathematics*. SIAM, Philadelphia, PA, USA, 1994.

H. Ohlsson, F. Gustafsson, L. Ljung, and S. Boyd. State smoothing by sum-of-norms regularization. *Automatica (to appear)*, 2011.

B. Oksendal. *Stochastic Differential Equations*. Springer, sixth edition, 2005.

J.A. Palmer, D.P. Wipf, K. Kreutz-Delgado, and B.D. Rao. Variational em algorithms for non-gaussian latent variable models. In *Proc. of NIPS*, 2006.

G. Pillonetto and B.M. Bell. Bayes and empirical Bayes semi-blind deconvolution using eigenfunctions of a prior covariance. *Automatica*, 43(10):1698–1712, 2007.

J. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods: support vector learning*, 1998.

M. Pontil and A. Verri. Properties of support vector machines. *Neural Computation*, 10:955–974, 1998.

C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

R.T. Rockafellar. *Convex Analysis*. Priceton Landmarks in Mathematics. Princeton University Press, 1970.

R.T. Rockafellar and R.J.B. Wets. *Variational Analysis*, volume 317. Springer, 1998.

S. Roweis and Z. Ghahramani. A unifying review of linear gaussian models. *Neural Computation*, 11:305–345, 1999.

S. Saitoh. *Theory of reproducing kernels and its applications*. Longman, 1988.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. (Adaptive Computation and Machine Learning). The MIT Press, 2001.

B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. *Neural Networks and Computational Learning Theory*, 81:416–426, 2001.

A. J. Smola and B. Schölkopf. Bayesian kernel methods. In S. Mendelson and A. J. Smola, editors, *Machine Learning, Proceedings of the Summer School, Australian National University*, pages 65–117, Berlin, Germany, 2003. Springer-Verlag.

R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B.*, 58:267–288, 1996.

M. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

P. Tseng and S. Yun. A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. *Comput. Optim. Appl.*, 47(2):1 –28, 2008.

V. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, USA, 1998.

G. Wahba. *Spline models for observational data*. SIAM, Philadelphia, 1990.

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and randomized GACV. Technical Report 984, Department of Statistics, University of Wisconsin, 1998.

D.P. Wipf, B.D. Rao, and S. Nagarajan. Latent variable bayesian models for promoting sparsity. *IEEE Transactions on Information Theory (to appear)*, 2011.

S.J. Wright. *Primal-dual interior-point methods*. Siam, Englewood Cliffs, N.J., USA, 1997.

K. Zhang and J.T. Kwok. Clustered nystrom method for large scale manifold learning and dimension reduction. *IEEE Transactions on Neural Networks*, 21(10):1576 –1587, 2010.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.